Bridging Chasms in Hindustani Music Retrieval

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

by

Joe Cheri Ross
(Roll No. 114050001)

Under the guidance of
Prof. Preeti Rao
and
Prof. Pushpak Bhattacharyya

Department of Computer Science & Engineering,
INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
December 2017

Declaration

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Joe Cheri Ross
(Roll No: 114050001)

Date: 22nd December 2017

Abstract

Development of a music recommender system, one of the key applications of Music Information Retrieval (MIR), necessitates research into methods to represent and retrieve music information efficiently. Considering the specific characteristics of each music culture and the diverse requirements thereof, the methods must be culture-aware with many of the associated tasks being culture-specific. This research is motivated by the importance of a music recommendation system for Hindustani music. The investigated tasks focus primarily on information extraction from melodic audio and text content, realizing the significance of information from multi-modal sources. In the context of information extraction from audio signals, we present our investigations on melodic motif detection involving *mukhda* (main title phrase of a composition) and *pakad* (raga characteristic phrase) detection. Our investigation on extracting meta-information from natural language text focuses on coreference resolution, with the aim of improving relation extraction from textual content. The task of raga similarity detection is investigated with both text and music content (available as music notation).

Melodic motifs form essential building blocks in Indian Classical music. The *mukhda* and the *pakad* phrases provide strong cues to the identity of the composition and the underlying raga respectively. Automatic detection of such recurring basic melodic phrases is highly relevant to music information retrieval in Hindustani music. This thesis discusses approaches to detect *mukhda* and *pakad* in a concert audio recording by exploring musicological cues and similarity computations.

Considering the large number of ragas in Hindustani music, detection of similarities between them is beneficial to music recommendation. The problem of raga similarity detection is investigated with two diverse data sources viz., discussions on Hindustani ragas and composition notations. Each of these sources help in extracting different aspects of raga similarity. While text discussions aid to extraction of similarities generally perceived by musicians, similarities based on melodic attributes are extracted through composition notations. Both the

approaches learn representations for ragas, and the similarities between the representations indicate the similarities between the ragas.

Realizing the importance of coreference resolution to improve relation extraction from text, our investigations on extracting meta-information focuses on the same from music discussion forums. The attempt to design a specific approach is motivated by the nature of the text and the domain specificity. While the feature design considers domain specificity and nature of the text, we also observe the need for a hybrid approach. The proposed modification to best-first clustering, for the clustering step in the mention-pair model, considers relation between candidate antecedents while resolving for an anaphoric mention. We also discuss a method to identify the semantic class, a crucial feature for coreference resolution, with the help of web resources. This approach to semantic class identification is generalizable for any domain-specific dataset with similar challenges. The investigations with eye-tracking and memory networks initiate research in the direction of bridging the gap between the cognitive process involved and the machine understanding of coreference resolution.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Music expresses that which cannot be said and on which it is impossible to be silent.

Victor Hugo

Music is an indispensable part of human civilization, and is probably the first cultural or artistic engagement known to mankind. On a very high level, music can be defined as sound organized in a time scale for favourable listening experience expressed vocally and through instruments. The relation between humans and music is not a recent development. No human culture on earth is an exception in the matter of music involvement, and the history of music in these cultures extend back to 250,000 years or more (Bannan, 2012). Music befitting the occasion is an integral part of any human gathering. Various studies have identified the role of music in developing societal bonding in human beings. With the advent of technology to store music, personalized music listening gained attention. This development is accompanied by methods to retrieve information from music with the aim to enrich music experience and cater to the requirements of certain associated fields. Towards an improved listening experience in any genre, personalized recommendation of music is indispensable. Similarity between performances is key to any music recommendation system. A majority of the Music Information Retrieval (MIR) tasks are directly or indirectly targeted to derive similarities between songs. There has been an extensive research in the past on MIR for Western music. This thesis presents our research directed towards a recommendation system for Hindustani music, considering the need of culture-specific approaches for this genre.

1.1 Music Information Retrieval (MIR)

With the advent of modern technologies, music has turned digital. Traditional ways of listening to music are replaced by personalized ways where the listener wants to enjoy what is most interesting to him/her (Casey et al., 2008). We have witnessed several revolutionary changes in the creation, storage, dissemination and listening of music during the last few years. The technological advancement has contributed to easy storage and access of the available content. Despite the amount of content available, the access to desirable content depends on the competence of the available music information retrieval methods. Taking this into account, development of tools for accessing, filtering, classification, and retrieving call for significant attention (Orio et al., 2006). Music information retrieval refers to extraction of information on genre, artist, repeating patterns, scores, lyrics from music etc. Research on music information retrieval brings together know-how of diverse areas including information technology, music theory, musicology, audio engineering, digital signal processing and cognitive science (Futrelle and Downie, 2002).

There exist approaches for content based MIR for music scores and audio data (Typke et al., 2005). Apart from the extraction from music content, it is also important to extract meta content available in the form of text to satisfy the requirements of MIR. Music data is not confined to audio recordings, but also includes symbolic notation, meta-data, artist information, lyrics and user-context information spanning across various modalities (Serra et al., 2013). While the musical characteristics are derived from audio recordings, the information extracted from musical experiences of people is complementary, but equally relevant. Apart from the available printed text on music, the vast range of websites, blogs and discussion forums are rich sources of music information (Serra et al., 2013). Having the associated meta information is important when a user searches for music content through a query containing meta information. The vast online text resources with rich music information are available in unstructured and semi-structured form. Representation of structured information extracted from both text and music content represented as a knowledge graph is helpful to MIR systems for better retrieval.

Most of the existing researches in MIR focus on Western music. Many classical music traditions including Indian, Chinese, Turkish, Arabian music etc. have rich musicological and cultural studies. Western commercial music of the past few decades has shaped a big part of the MIR related research, giving little attention to information extraction tasks in other cultures (Serra, 2011). The current MIR methodologies do not work for many problems in these tradi-

tions and many times the MIR requirements are also different (Serra, 2011). For instance, raga recognition problem is very specific to Indian classical music.

1.2 Introduction to Hindustani Music Concepts

The North Indian tradition of Indian classical music is referred to as Hindustani music (the South Indian tradition is Carnatic music). It has been evolving since the 12th century in North India with significant influence from Persian and Arabian music (Junius et al., 1969; Wikipedia, 2017a). Classical music traditions of Afghanistan, Pakistan, Nepal and Bangladesh also exhibit similarities to Hindustani music. Indian music is essentially melodic and monophonic (or more accurately, heterophonic) in nature. In spite of this apparent simplicity, it is considered a highly evolved and sophisticated tradition. In this section, we introduce Hindustani music and the main concepts which are essential to better understanding of this thesis. The majority of the content in this section is taken from the books Shruti (Bagchee, 2006), Nad: Understanding Raga Music (Bagchee, 1998) and from the paper by Rao and Rao (2014a).

1.2.1 *Bandish* and Performance

In Hindustani music, a composition is referred with the term *bandish*. Generally *bandishes* are very short with 2 parts; *sthayi* and *antara*. For the artiste, a *bandish* is a means to present the raga, and the words of the *bandish* are given less importance compared to the raga of the composition. A composed musical piece termed as *bandish* is written to perform in a particular raga, giving ample freedom to the performer to improvise upon. As the literal meaning suggests, *bandish* is tied to its raga, *tala* (rhythm) and lyrics. *Bandish* is taken as the basic framework for a performance, which gets enriched with improvisation while the performer renders it. Realization of a *bandish* in a performance brings out all the colors and characteristics of a raga.

In a typical Hindustani performance, the main artiste (a vocalist or an instrumentalist) renders the melody. The main artiste is accompanied by a percussionist on Tabla and an accompanist on a stringed instrument or harmonium provides the secondary melody (Rao and Rao, 2014a; Bagchee, 2006). The artiste first starts with *avachar* introducing the basic melodic pattern of the raga. This is followed by commencement of *sthayi*, called as *alap*. *Alap* is sung with the permitted raga notes to the vowel 'a' (*alap* in *akar*) or with the words of the *bandish*

(*bol-alap*). Tabla joins the performance after *alap* when the presentation of composition starts. Throughout the performance, the words from the lyrics are repeated with different emphasis and ornamentation. Each *bandish* is identified by its main title phrase called as *mukhda* and the artiste often returns to this phrase in a performance. This recurring phrase has a fixed melodic shape and the last syllable of the *mukhda* coincides with the emphatic beat of the rhythmic (*tala*) cycle called *sam*. The fast passages towards the end of a performance are referred to as *tan*. *akar tan* is sung with the vowel 'a' and *sargam tan* uses note names.

Figure 1.1: A Hindustani vocal concert setting. A vocalist accompanied by a harmonium player, tabla player and two tanpura accompanists.

1.2.2 Swara

The term *swara* refers to note in both the classical music traditions. The seven main *swaras* are known as *Shadja* (*Sa*), *Rishab* (Re) , *Gandhar* (*Ga*), *Madhyam* (*Ma*), *Pancham* (*Pa*), *Dhaivat* (*Dha*) and *Nishad* (*Ni*). The note solfege is given within parenthesis. The set of seven *swaras* is called *saptak* and the eighth note is the repetition of the first note (with double the frequency). These natural notes are referred to as *shuddha swaras* and the flat or sharp versions of some of these notes are called *vikrit swaras*. A *komal swara* denotes the flat version of a *swara* and a *thivra swara* denotes the sharp version of a *swara*. Except *Sa* and *Pa*, all *swaras* have a *komal* or *thivra*. Figure 1.2 shows the notes correspondence to keys in a keyboard when the tonic (*Sa*) is at C.

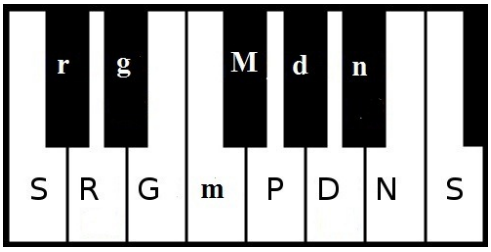


Figure 1.2: *Swara* markings on a keyboard. source: (Suja, 2012)

Table 1.1 gives a description of Hindustani *swaras* with details on solfege and the corresponding western solfege. The notation id mentioned in this table is as per Bhatkande notation system (using Latin script). We use this notation scheme throughout this thesis.

Note name	Notation ID	Solfa syllable	Full name	Western Solfa
sa	S	sa	Shadja	doh
re (komal)	r	re	Rishabha (komal)	
re	R	re	Rishabha (shuddha)	re
ga (komal)	g	ga	Gandhara (komal)	
ga	G	ga	Gandhara (shuddha)	mi
ma	m	ma	madhyama (shuddha)	fa
ma (thivra)	M	ma	madhyama (thivra)	
pa	P	pa	panchama	sol
dha (komal)	d	dha	dhaivata (komal)	
dha	D	dha	dhaivata (shuddha)	la
ni (komal)	n	ni	nishada (komal)	
ni	N	ni	nishada (shuddha)	ti
sa	S’	sa	Shadja	doh

Table 1.1: Detailed description of Hindustani *swaras* (Sadhana, 2011; Bagchee, 1998)

1.2.3 Raga

The compositions and their performances in both these classical traditions are strictly based on the grammar prescribed by the raga framework. A raga is a melodic mode or tonal matrix providing the grammar for the notes and melodic phrases, but not limiting the improvisatory

possibilities in a performance (Rao and Rao, 2014a). Raga is the foremost, and possibly the most rudimentary and fundamental concept in both the classical music traditions of India. In principle, a raga (the closest concept in Western music is the mode) is a melodic framework where certain rules are applicable to a set of notes. A raga uses only certain notes which are permitted by the raga rules and the allowable ascending sequence of the notes is called *aroha* (ascent) and descending sequence of the notes is called *avaroha* (descent). Five natural notes and their sharp and flat forms constitutes a scale, and a selection of seven notes from these twelve notes decides the parent scale of a raga. The parent scale of a raga is referred to as a *thaat*. With a few exceptions, all ragas have atleast 5 notes. For instance, the *thaat* with all *shuddha swaras* is *Bilawal*. A raga should have the tonic *S*, and either *M* or *P* must also be present.

The major difference between the ragas within a *thaat* lies in the way in which the permitted notes are used. While the identity of certain ragas depend on the dominant notes which are stressed by their frequency or by the duration they are played, the way in which certain notes are intoned or ornamented contributes to the identity of others. The tonic and the *vadi* (sonant) are considered to be the strong notes in a raga and most of the variation begin and end at *vadi*. The fifth note from *vadi* is referred to as *samvadi* (consonant).

The characteristic melodic phrases are of prime importance to the identity of a raga, and they are considered to be the signature of a raga. In a raga performance, these phrases are repeated and it is key to identification of the raga by the listener. Characteristic phrase of a raga is also referred to as *pakad*. In the performance of a *bandish* in a raga, the artiste enriches the performances with his/her interpretations and improvisation adhering to the mentioned conventions.

1.2.4 *Tala* (Rhythm)

Along with melody and harmony, rhythm is also one of the fundamental elements of music. The time arrangement in music is referred to as rhythm. In Western music, metre is the rhythmic structure (Wikipedia, 2017b), indicating the beat pattern. A measure in written music divides the beats into smaller groups according to the metre of the music. The term corresponding to metre in Hindustani music is *tala*, and to measure is *vibhaga*. For instance, *Dadra tal* cycle comprises of a 3+3 patterns with 2 *vibhagas*, with each *vibhaga* having 3 beats. *Jhap tal* having a 2+3+2+3 pattern is considered to be more complex with 4 *vibhagas*. First and third *vibhagas*

have 2 beats, and second and fourth *vibhagas* have 3 beats each. There are 3 kinds of beats: the emphatic beat known as *sam*, empty beat known as *khali* and others are called *tali*. The emphasis on the beats are shown as in this example of *Dadra tal*.

$|1_X 23|1_O 23|$

The first beat or *matra* of the first section is marked with an emphasis, whereas the the first beat of the second section is marked with lesser emphasis. When played by a drummer, generally hard stroke indicates a *sam* and soft stroke indicates a *khali*. The term *laya* denotes the speed or tempo in which a performance proceeds. Widely used *layas* are *vilambit* (slow), *madhya* (medium speed) and *drut* (fast).

Tabla, the most important percussion instrument in Hindustani music was introduced by Amir Khusro in the 13th century. This has two drums distinguished by their shapes; *bayan* and *dayan*. *Bol* denotes the syllables corresponding to the beats in a *tal* cycle. These syllables correspond to various strokes in a percussion instrument.

1.2.5 *Mukhda* and *pakad*

Hindustani music, especially the modern khyal style, is a predominantly improvised music tradition operating within a well-defined raga (melodic) and tala (rhythmic) framework. While the main title phrase *mukhda* is central to the identity of a *bandish,* the characteristic phrase or *pakad* reveals the identity of the raga a *bandish* is based on. As mentioned before, *mukhda* rendition is associated with the emphatic stroke *sam*. In the improvised section of the concert known as the *bol-alap*, the singer elaborates within each rhythmic cycle of the *tala* using the words of the *bandish* interspersed with solfege and held vowels, purposefully reaching the strongly accented first beat (the *sam*) of the next rhythmic cycle on a fixed syllable of the signature phrase of the *bandish*. It acts like a refrain throughout the exposition, which can last several minutes, whereas the other lyrics of the *bandish* can undergo extensive variation in melodic shape in the course of improvisation.

Apart from the permitted scale intervals (*swaras*) that define a raga, it is its characteristic phrases that complete the grammar and give it a unique identity (Rao et al., 1999). "While a singer moves through the raga (*calna*) to understand it, a listener attempts to catch the motion (*pakad*) to identify the raga at hand" (Rahaim, 2012). Most ragas can be adequately represented by up to 8 phrases which then become the essential building blocks of any melody. For example, *P M G m G* is a characteristic phrase of raga *Bihag* and *N R G M* is the character-

istic phrase of raga *Yaman*. Thus the detection of recurring phrases can help to identify the raga. Indeed, musical training involves learning to associate characteristic phrases, or motifs, with ragas. Melodic improvisation involves weaving together a unique melody bound by the chosen rhythmic cycle (*tala*) and consistent with the raga phraseology. A characteristic phrase is defined by the sequence of *swaras* involved, their relative duration and ornamentation. It is observed that the melodic shape of a characteristic phrase change with speed. There are a few ragas with no *pakad* and their identity rely on the modal characteristics.

1.3 Music information retrieval for Hindustani music

As mentioned, there is a need for culture-aware approaches for music information retrieval in Hindustani music, considering the problems and challenges in this genre of music. Most of the identified tasks are common to both the traditions in Indian music. The concept of raga in Indian classical music defines a melodic framework for the compositions in Hindustani and Carnatic, whereas the closest concept in Western music is the mode. Mode is comparable with the concept raga only with respect to the scale. But the grammar of a raga is even stricter, with rules on allowed phrases and transition between notes. Indian classical music follows a vocal tradition with vocalist as the centre of attention in a performance and the instruments provide accompaniment to the singing. Western music gives more importance to harmony, arranging multiple instruments and chords in its compositions. Just these differences alone introduce a set of tasks in Hindustani music which are not relevant to Western music.

Taking into account the significance of raga to Hindustani music, raga based classification and retrieval of performances is of prime importance to Hindustani MIR. Automatic raga identification is one of the most researched problems in Hindustani MIR (Chordia and Rae, 2007; Pandey et al., 2003; Belle et al., 2009; Kumar et al., 2014). Apart from its relevance to MIR, raga identification throws light on how implicit music knowledge is manifested in a performance (Rao and Rao, 2014a). Detecting similarities between ragas is crucial to music recommendation tasks in Hindustani to recommend performances in a related raga. There exists a large number of ragas in both the tradition and all aspects of similarities between them are not available in documented form. This is little explored and this is one of the researched problems in this thesis. Likewise, identifying similarities between artistes, compositions etc. have similar applications.

Many of the well studied computational tasks in Hindustani music are fundamental problems, which are building blocks to other high level tasks. In Indian music, tonic is chosen by the performer and need not always be an absolute pitch. For any study involving analysis of melody, identification of tonic is important (Rao and Rao, 2014a). There are a few attempts to solve this problem (Gulati et al., 2012, 2014; Bellur et al., 2012b; Salamon et al., 2012). As in many other music cultures, segmentation and labeling of motifs in a performance is a relevant problem in Hindustani music. This problem is in its infancy. We will be discussing methods for detection of *mukhda* and *pakad* instances from a Hindustani performance.

Considering the strong binding of the metrical structure with melodic aspects of a performance, estimation of different aspects of metre is an important MIR task. Computational modeling of rhythmic structure can help tasks like automatic segmentation, which is an essential prior step to motif identification and structural analysis tasks. These elementary rhythm based tasks include detection of note onsets, tempo, beats and downbeats. *Alap* segmentation task discussed by Vinutha and Rao (2014) takes clues from significant changes in rhythmic structure in a performance . Along the similar lines, TP et al. (2016) investigates structural analysis of metered section (*gat*) of sitar and sarod concerts .

While many researches for Western music rely on symbol scores as the data source, Hindustani music prefer audio performances. Symbolic notation available for Hindustani does not encapsulate the details of how the composition is to be performed. Symbolic notation available for Hindustani music only provides an abstract skeleton for the performer to improvise, and cannot be considered as the complete representation of a performance. The ornamentations and intonations found in a performance are crucial for many MIR tasks. So far, no successful attempts were reported to symbolically represent a Hindustani performance. These considerations restricted the Hindustani MIR research to rely on audio content of the performances as the data source. The tasks for which melodic information is essential, pitch estimation methods are helpful to extract melodic information in the form of pitch sequence. There are a few approaches to extract predominant-F0 from audio content even in the presence of percussive and pitched accompaniment (Rao and Rao, 2010; Salamon and Gómez, 2012).

1.4 Knowledge representation: MIR

Identifying the right knowledge representation methods is crucial to any knowledge based artificial intelligence problems. It helps to formally represent the available information in the domain of application. Even though this thesis does not address this aspect of MIR, this section gives a brief introduction to the necessity and approaches for knowledge representation. The efficiency of the inference system which draws conclusions based on the possessed knowledge is dependent on the expressibility of the knowledge representation. The implicit information from the knowledge base is inferred from the explicit information represented in the knowledge base. Information retrieval in any domain having vast amount of information need to have a formalized information representation. Semantic web technologies help to build an ontology which defines the concepts and relations between the concepts in a domain.

To facilitate better music information retrieval, integration of knowledge representation methods is essential to connect meta data, content based music information and musicological information. The music information distributed across different sources can be connected if represented with standard knowledge representation methods, thus creating a unified information source. Available semantic web technologies aim at formalizing knowledge representation and provides with tools for accessing web information and inferring from existing knowledge. Semantic web is in progress, interconnecting the variety of information making them better accessible.

There has been different approaches to improve MIR through knowledge representation. One of the prominent ontologies in music domain is the music ontology proposed by Raimond et al. (2007). This ontology is designed primarily for representing editorial information, musicological information with information on content of music item and workflows. Though it has capabilities to contain information about the musical content, it is not sufficient enough to represent content based information pertaining to many music cultures. Extension of existing music ontology with capabilities to represent raga information, details of detected motifs etc. will help better retrieval of Indian music with attributes relevant to Indian music. For Carnatic music, Gopal and Serra has proposed raga ontology and its semantic substructures including *swara*, *gamaka*, phrases and *tala* ontology (Koduri, 2016; Koduri and Serra, 2013).

1.5 Motivation

Music recommendation is a commercially important area of study in MIR, which manifests as a natural use case for almost any information extracted from different modalities. With the increasing amount of available music content, personalized recommendation of songs matching the taste of a user is a complex challenge. Aesthetic preferences of a user is quite subjective, and any logical recommendation can only be made by analyzing the different aspects of similarities with the songs he/she listens to. Considering the listening scenario of Hindustani music into account, personalized ordering and recommendation of music performances demands further research. The existing music recommendation systems are realized with content-based filtering, context-based filtering, collaborative filtering or a hybrid method comprising combination of others. Content-based and context-based techniques utilizes information extracted from content and meta-data to identify song similarity. Collaborative filtering considers listening trend of other users and the relation between the users, to recommend songs (Celma, 2010).

This thesis focus on information extraction tasks which are relevant to the attributes of similarity for Hindustani classical music. Locus of this inquiry is about the relevant similarity elements in recommending relevant songs to the user and how these similarities can be extracted, given the user preferences. Even when content based similarities form the core of the recommendation system, meta-information also plays a significant role in recommending relevant performances. For any music genre, details including performing artiste, composer, etc. can only be obtained from the meta-data available from cover arts, books, online articles and discussion forums.

In Hindustani music, content based similarities heavily depends on the raga based similarities. Considering the importance of raga based grouping of performances for recommendation, raga recognition is an important task. The relevance of *pakads* to the identity of a raga in a performance, makes detection of *pakad* instances crucial to raga recognition. Ganguli and Rao (2017) shows the importance of a method based on phrase based similarity to identify distinction between similar ragas, compared to methods based on pitch distribution. Also, *pakad* is an excellent means to identify songs with stronger similarity within the same raga, facilitating relevant recommendations within the same raga. For instance, within a raga, certain performances may have a *pakad* occurring more prominently than in others. The user listening to one of them, may anticipate another one from the same group. These motivations led to our investigations on

melodic motif detection from Hindustani audio performances, dealing with *pakad* (characteristic phrase) and *mukhda* (title phrase). Even though *mukhda* detection has more pedagogical applications, it is also important to identification of similar songs beyond the constraints of raga. Since *mukhda* is considered as the signature phrase of any composition, the similarity between compositions based on the *mukhda* phrase is more meaningful than analyzing the similarity based on the whole compositions.

Besides recommending songs within the same raga, the user may also appreciate suggestions from related ragas. A user listening to a performance in raga Kalyan may appreciate a performance from raga Yaman as well. This requires the similarities between ragas to be extracted, which are not available documented. This motivated investigations to detect raga similarities from two diverse sources, involving melodic content and natural language text. The first approach extracts similarities based on melodic attributes from *bandish* notation data. The other approach depends on textual discussions on Hindustani music. Here, we intend to extract similarities which are generally accepted by musicians and more likely to reflect in these discussions.

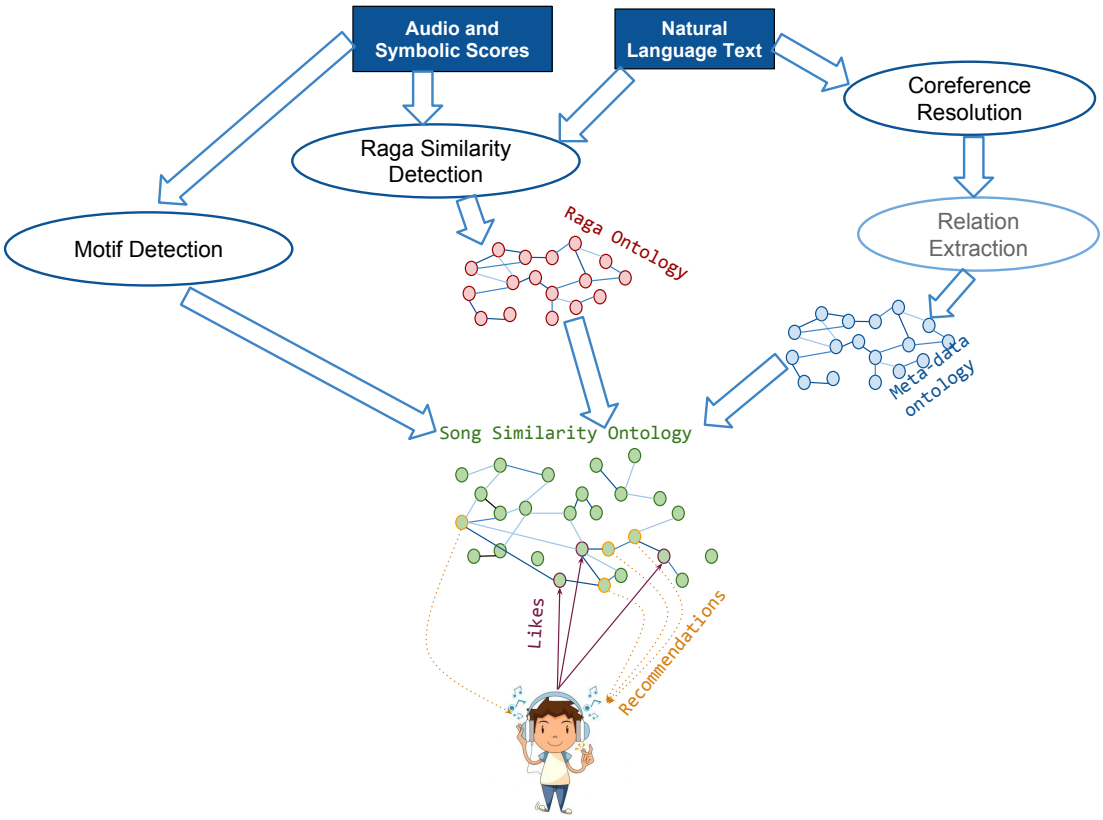


Figure 1.3: Selection of tasks to assist music recommendation in Hindustani music

As mentioned, information extracted from natural language text content is pertinent to a recommendation system. It is very likely that, a user may prefer songs from the same artiste or composer. Identifying the entities, and the relation between the entities from the text available as blogs, forums etc. are fundamental to acquiring this knowledge. Also, certain information which is generally obtained from audio content, can be extracted from text content as well (eg. raga of a composition), providing a fallback option. For this study, we consider Rasikas.org which is a popular and active discussion forum in Carnatic music. The availability of substantial text content in this discussion forum is the major motivation behind its selection. Also, the domain of Carnatic music is very similar to Hindustani music.

Before addressing the problem of relation extraction, we decided to solve a more fundamental problem; coreference resolution. Resolution of anaphoric noun phrases is imperative to better relation extraction from text. Consider the following forum post on raga *Priyadarshini* from Rasikas.org:-

Rag Priya Darsini is a janya of 27th mela Sarasangi. The aroh and avaro., are SRMDNS - SNDMRS. This rag is popular more through bhajans and occasionally I had heard in Katcheries too. The rag evokes bhakti. The pancham varjya and the stress on the note nishad adds beauty to this rag. Sri. Thanjavoor Thyagarajan had once given a very elaborated raga alapan of this rag for a Narayana teertha tharangam song. He is expert particularly for this rag.

In this forum post, the resolution of 'this rag' and 'the rag' can help in extraction of more information related to 'the raga Priyadarshini'. Resolution of 'He' in the last sentence helps in associating 'Thanjavoor Thyagarajan' with 'raga Priyadarshini'. This example shows the relevance of coreference resolution in improving the meta-information extraction from the available text content in these forums.

In spite of the fact that there exist quite a lot approaches for coreference resolution, this is investigated again considering the domain specificity and nature of the text. In a specific domain, the type of entities, the relation between the entities and the choice of words in the description differ from any other domain. For instance, distinguishing the entities as person, song, performance, instrument is important to the domain of music. The nature of the text is also a factor to be considered while designing the approach. Discussion forum posts are in general short discourses with informal text having quite a lot grammatical errors.

This thesis is a journey towards building a knowledge base with information required for a Hindustani recommendation system. Along with automatically extractable information, the

knowledge base also encompasses manually annotated information for completeness. Based on the aforesaid motivations, this thesis addresses three important problems for Hindustani music recommendation system viz., melodic motif detection, detection of raga similarity and coreference resolution. While the tasks of motif detection and the first approach for raga similarity detection relies on audio and melodic content, the other approach for raga similarity detection and information extraction from text depend on text content. Figure 1.3 depicts how the selected tasks are important to compute similarity between songs in order to make relevant recommendations.

1.6 Objectives

Here we briefly describe the objectives of the selected tasks.

- **Melodic motif detection**: This task is on detecting instances of a melodic motif from a Hindustani classical performance, given template(s) of the motif. This work focuses on detection of two kinds of motifs in Hindustani music; *mukhda* (main title phrase of a composition) and *pakad* (raga characteristic phrase). Comprehending the variabilities and invariabilities within instances of a motif is key to its detection from a performance.

- **Raga similarity detection**: This targets detection of similarities between ragas. This study aims to extract similarities from textual discussions as well as composition notation. The approach for textual discussions look for similarities generally perceived by musicians and cannot be derived explicitly from the raga attributes. Similarities solely based on melodic attributes are derivable from composition notation.

- **Coreference Resolution**: The task of coreference resolution resolves anaphoric mentions in the text with the entities it refer to. We investigate this problem specific to forum posts from Rasikas.org. The domain and the nature of text are the key considerations to the design of an approach.

1.7 Overview of Thesis Contribution

Here we give an overview on the contributions of this research discussed in the subsequent chapters.

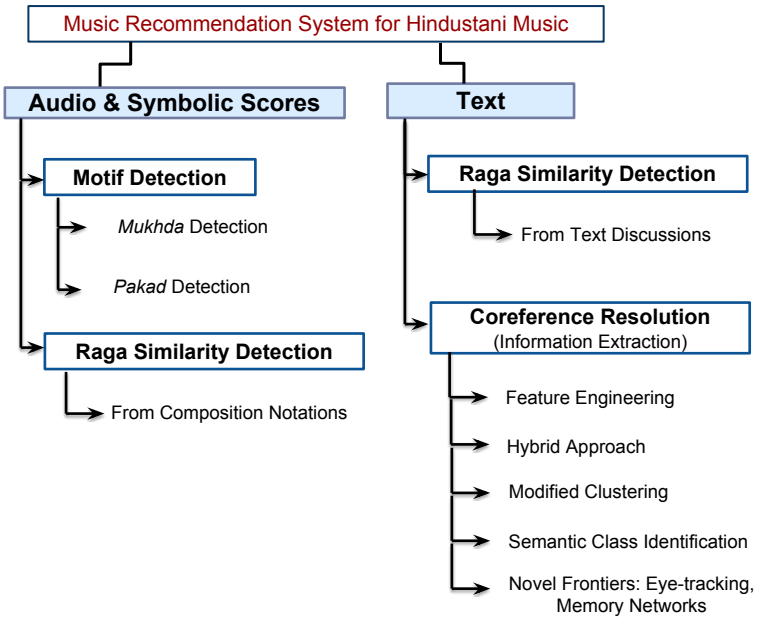


Figure 1.4: Thesis contributions

1.7.1 Melodic Motif Detection

Melodic motifs form essential building blocks in Indian classical music. The motifs, or key phrases, provide strong cues to the identity of the underlying raga in both Hindustani and Carnatic styles of Indian music. Thus the automatic detection of such recurring basic melodic shapes from audio is of relevance in music information retrieval. The extraction of melodic attributes from polyphonic audio and the variability inherent in the performance, which does not follow a predefined score, make the task particularly challenging. Despite the scope of improvisation, there are invariabilities associated with each class of motif. Comprehending these invariabilities and its range is key to identification of motif instances from a performance. Given the well-known difficulties with extracting low-level musical attributes such as pitch and onsets from general polyphonic audio recordings, most work in motivic analysis for music has been restricted to symbolic scores. For our tasks, we depend on a predominant pitch detection algo-

rithm to extract the melodic pitch contour from vocal concert recordings (Rao and Rao, 2010) and motifs are identified from the vocal pitch time series.

Mukhda Detection

Here we discuss a method to automatically identify *mukhda* instances in a *bandish* performance. The relation to the rhythmic cycle serves as a strong cue in this task. The occurrence of *mukhda* has association with the rhythmic cycle; *mukhda* is rendered in a way that a particular syllable of *mukhda* is always sung at the strongly accented down beat (*sam*) of the rhythmic cycle. In this work, we consider the segmentation of audio signals with the mentioned rhythmic cues and identifying the *mukhda* instances by computing similarity measures on time series of pitch values. For a particular performance, candidate phrases are identified from all the identified *sam* occurrences with the help of given information on the *mukhda* duration and its relation with a *sam* instance. For identifying *mukhda* phrases from these candidates, we experiment with two major similarity measures, Symbolic Aggregate approXimation (SAX) (Lin et al., 2003) and Dynamic Time Warping (DTW) (Berndt and Clifford, 1994) which are widely used for time series matching.

The experiments are performed on a selected set of 4 *bandish* performances of well-known Hindustani *khyal* vocalists. All *mukhda* instances are manually labeled which serve as the ground-truth for the experiments. To maximize the use of the available annotated data, each labeled motif is considered as the reference with all other motifs serving as positive tokens and the remaining candidates as negative tokens. The data representations chosen for the study are either the continuous pitch values (i.e. 1200 cents per octave) or the quantized versions (12 semitones per octave on an equi-tempered scale). The results show the effectiveness of this approach. The identification accuracy of DTW is better compared to SAX showing the need for non-uniform time warping.

Pakad Detection

The identity of a *pakad* lies in the notes combination, the intonation of the notes and the ornamentations. Preserving the characteristics of a *pakad*, the artiste is granted the flexibility to improvise it. Despite the variations present, the detection method is expected to discover instances of a *pakad* from a performance. In this work, we restrict ourselves to the problem of classifying pre-segmented phrases from the pool of phrases belonging to different *pakad*

classes having high inter-class similarity. The focus is more on the similarity measure to better distinguish between dissimilar *pakad* instances. For this study, we constructed a dataset of performances in raga Alhaiya-Bilawal having *pakads DnDP, mnDP, GRGP* [1] . One performance from raga Kafi is taken because it also has DnDP as a *pakad*, but with different characteristics. Considering the variabilities and invariabilities permitted in the rendition of a *pakad*, here we try to analyze how learned constraints can be better over Sakoe-Chiba constraints for DTW similarity measure. Constraints are learned from the *pakad* instances from the training data; a subset of the dataset.

We consider the retrieval of *DnDP* and *mnDP* phrases given reference templates from the training set of the same phrase classes. The test set includes all the phrases of the corresponding class drawn from the remaining concerts in the dataset, plus all the phrases those are not in that class across all the concerts. The retrieval performance for a given phrase class is measured by applying a threshold to the distribution of DTW distances obtained. We observe that retrieval performance of DTW with global constraints clearly surpasses that without constraint. The learned constraints provide for superior or similar hit-rate.

1.7.2 Raga Similarity Detection

Raga being one of the most prominent categorization aspect of Hindustani music, identifying similarities between them is of prime importance to many Hindustani music specific tasks such as music information retrieval, music recommendation and automatic analysis of large-scale musical content *etc*. Generally similarity between ragas is inferred through attributes associated with the ragas. For instance, in Hindustani music, classification of ragas based on the tonal material involved is termed as *thaat*[2]. The similarities between ragas is not always a function of their attribute-wise similarities. For instance, *Darbari Kanada (asavari thaat)* and *Kaunsi Kanada (bhairavi thaat)* are considered similar inspite of their different *thaats*. Here we attempt to extract similarities between ragas with the help of learned representations of the ragas. The similarity (say cosine similarity) between the representation gives the similarity between the ragas. We investigate on learning representations from two diverse sources; text discussions on Hindustani ragas and composition notations.

[1] In Hindustani notation system S r R g G m M P d D n N corresponds to C $C^{\#}$ D $D^{\#}$ E F $F^{\#}$ G $G^{\#}$ A $A^{\#}$ B notes in western music notation system, when the tonic is at C

[2] `https://en.wikipedia.org/wiki/That_(music)`

From text discussion, we intend to extract similarities which are generally accepted by musicians and that are likely to appear in text discussions. Similarities based on melodic attributes are extracted from composition notations. The representation learned from text discussions are word vectors for the raga words. Here we propose an incremental approach modifying Mikolov's approach (Mikolov et al., 2013a) for extracting word vectors from a relatively smaller corpus. The non-availability of ground truth for this task makes the evaluation challenging. We designed a quantitative evaluation method considering one strong attribute of similarity between ragas, the *thaat*. Also with the help of a trained Hindustani musician we checked for the effectiveness of this method by identifying clusters of similar ragas from the t-SNE visualization of the raga word vectors. Quantitative evaluation shows the effectiveness of the proposed method against the baselines designed with Mikolov's standard approach. Also the musician's observations confirm the validity of this approach.

The approach for composition notation learns embeddings for notes for each raga. We hypothesize that embeddings learned for a given note for similar ragas will have more similarities. For example, the representation for note *Ma-elevated* (equivalent note *F#* in C-scale) in raga *Yaman* can be expected to be very similar to that of *Yaman Kalyan* as both of these ragas share very similar melodic characteristics. We design a deep recurrent neural network (RNN), with bi-directional LSTMs as recurrent units, that learns to predict the forth-coming notes that are highly likely to appear in a *bandish* composition, given input sequences of notes. This is analogous to neural language models built for speech and text synthesis (Mikolov et al., 2011). The embedding layer of the network learns the note-embeddings for a raga. We rely on two different evaluation methods to validate our approach. The first one is based on perplexity that evaluates how well a note-sequence generator model (neural-network based, n-gram based *etc.*) can predict a new sequences in a raga. Since note-embeddings are an integral part of our architecture, a low-perplexed note-sequence generator model should learn more accurate note embeddings. The second method relies on clustering of ragas based on different raga-similarity measures computed using our approach and baselines. We compare our approach with baselines based on n-gram profile, pitch-class distribution and a neural network similar to the proposed design, with uni-directional LSTM instead of bi-directional. The evaluation results show the superior performance of the proposed approach over the baselines. Clustering results show that the performance of uni-directional LSTM based approach is comparable with that of bi-directional.

1.7.3 Coreference Resolution

In the context of MIR, information extraction from text helps in extracting meta information from the large available online content to enrich the music information knowledge base along with information extracted from audio content. Some discussion forums and blogs are rich sources of information pertaining to Indian classical music. We discuss the task of coreference resolution with the aim to improve knowledge extraction from discussion forums on Indian classical music. Even though coreference resolution is a widely researched problem, we investigate on this problem considering the nature of the text in discussion forum and the domain. As mentioned, we consider Rasikas.org, a Carnatic music discussion forum due of its popularity and the amount of content available. Each forum post is considered as a document having an average size of 7 sentences.

Feature Engineering and Hybrid Approach

We follow a supervised learning approach, formulating the problem using mention-pair model, where the coreference classification decision is made for the mention pairs. The classification is followed by clustering, forming distinct coreferent chains (Recasens and Hovy, 2009; Aone and Bennett, 1995; McCarthy and Lehnert, 1995). During the mention-pair classification step, features are computed for a mention pair comprising of an anaphoric mention and a candidate antecedent mention. We make use of a subset of conventional features including the features described in Soon et al. (2001). Apart from proposing dataset-specific modifications to some of the conventional features, we also propose a few novel features mostly derived from grammatical role (extracted from dependency parse information) based features. We experimented with different classifiers for mention-pair classification and best-first clustering (Ng and Cardie, 2002b; Aone and Bennett, 1995) for the clustering step. The feature evaluation experiment to identify the relative importance of features affirms our hypothesis that, for short documents grammatical role features are more important than for longer documents.

In the initial phase of this work, we had only a limited number of coreference annotated forum posts. To deal with data insufficiency, we tried with Bayesian network with hand-engineered network structure for mention pair classification. Antal et al. (2004) claims that, when the data is scarce and we know the dependencies between features, Bayesian network performs better. The results showing better accuracy with Bayesian network on this smaller

dataset supports this claim. The human process of coreference resolution involves heterogeneous classification tasks depending on the type of mention which is resolved. The knowledge applied for resolving a proper noun anaphoric mention is different from knowledge applied for resolving a pronominal anaphoric mention. Further investigation led to a hybrid approach combining rule-based approach and machine learning (ML) based approach, segregating the classification of mention pairs which are better classified by rules. The results show that, across all the classifiers hybrid approach has a better accuracy over the pure machine learning based approach.

Improved Best-First Clustering

Best-first clustering is used so far for our experiments to form coreferent chains from mention pairs classified as coreferent. Best-first clustering selects the antecedent from the mention pair having the highest classification confidence score associated with it. The probability estimate of mention-pair classification is used as the confidence score. The existing Best-first clustering method considers only the relation between the anaphoric mention and the candidate antecedents to select the best antecedent. This proposed method considers the relation between the candidate antecedents, along with the relation between the anaphoric mention and the candidate antecedents. This is motivated by the fact that, if there is a stronger coreferent relation between two candidate antecedents, one of them is more likely to be the best antecedent compared to other candidate antecedents. Based on this, we compute a score for each candidate antecedent referred to as 'support-score'. Then the classification confidence of the mention pair involving this candidate antecedent and the anaphoric mention is recomputed as a linear combination of the probability estimate (previous classification confidence) and the support-score associated with the candidate antecedent.

The basic version of this modified clustering employs constant values for linear combination coefficients. We also experiment with 2 variants of the same approach, where the linear combination coefficient is decided dynamically. With the proposed approach, we observe a modest but statistically significant improvement over the best-first clustering (baseline) for this dataset. The variants of this approach gives improvement over the baseline, but not over the improved approach.

Semantic Class Identification

Understanding the semantics play a crucial role in coreference resolution. Semantic class (named entity class) compatibility between the entity mentions is imperative to coreference decision. In generic domain data, the semantic classes include person, location, organization and geo-political entity (GPE) etc. But in this domain, *person, raga, song, music instrument and music concept* are the relevant semantic classes. The aim of this task of semantic class identification is to classify a named entity into one out of the defined semantic classes. As the case with our dataset, short and noisy text containing the entity makes it challenging to extract the semantic class of the entity through the context. Also the non-availability of a named-entity annotated training dataset for this content adds to the challenge. We introduce a method for semantic class identification for a given entity using the web when the entity boundaries are known. We propose this as a generalized method applicable to any specific domain. In case of such specific domains, there may be several challenges: (i) There may be insufficient or no annotated data available for training a named entity recognition system (ii) The text may be noisy making it difficult to consider context, or (iii) the named entity classes include domain-specific classes. We present three approaches to use the web: (a) a baseline, rule-based approach that uses a structured web repository (Freebase), (b) a supervised approach that uses search engine results and topic models, and (c) a supervised approach that improves upon (b) with task-specific hierarchy of classifiers. The methods (b) and (c) depend on the web to gather context for a given entity, to determine its semantic class. We used the proposed method for semantic class identification for all the mentions in Rasikas.org dataset. This is utilized for the semantic class compatibility feature, resulting in improvement of coreference resolution accuracy.

New Frontiers in Coreference Resolution

The task of coreference resolution involves pragmatics and world knowledge. Though there exist many approaches for coreference resolution, they are way behind imitating the human process of coreference resolution. These traditional systems are yet to achieve acceptable accuracy and are heading towards a saturation barrier. This is an initiative in a different direction to bridge the gap between human cognitive process and machine understanding of coreference resolution. This work discusses two different attempts to incorporate ideas from human cognition for coreference resolution, using eye-tracking input and memory networks (Weston et al.,

2014). The former approach uses information from human cognition through gaze data from eye tracking, collected during coreference annotation task by the participants. The latter intends to simulate human cognition for coreference resolution through memory networks. We avoid the use of our dataset in these experiments. A subset of MUC dataset (a general domain dataset) is chosen for eye-tracking experiments to avoid the bias that a domain-specific dataset may induce. A synthetic dataset is created for memory networks experiments in order to reduce the adverse effect of noise in real-world data.

The task with eye-tracking aims at utilizing cognitive information obtained from the eye movements behavior of annotators for automatic coreference resolution. In the early phase of this initiative, to get rid of the noise while capturing cognitive information pertaining to coreference resolution from reading experiments, the readers for the experiments were asked to do coreference annotation. We first record eye-movement behavior of multiple annotators resolving coreferences in 22 documents selected from the MUC dataset. We observe that gaze transition probability derived from regression counts associated with a mention, signify the candidacy of that mention as an antecedent. We experiment with a supervised system following a mention-pair model (Soon et al., 2001)- injecting the eye-movement information into it. Eye tracking information is utilized in the process of mention pair pruning prior to mention pair classification. Our heuristic brings noticeable improvement in accuracy with different classifiers.

Human process of coreference resolution involves comprehending a discourse, along with using information from other knowledge sources like lexical, syntactic and semantic information. The comprehension inability of most of the existing approaches restrict them resolving coreference mentions which require context information and holistic understanding of the discourse. The potential of memory networks (Weston et al., 2014) towards comprehending the context of a discourse motivates this initiative. Memory networks integrate a memory component and inference capability which are jointly used to comprehend a discourse and perform reasoning (Weston et al., 2014; Sukhbaatar et al., 2015; Kumar et al., 2015). We utilize memory networks for coreference resolution, modeling it as a question answering task. The question is on the anaphoric mention in the document and the answer is its antecedent. For our experiments we create a set of synthetic data with varying difficulty levels. End-to-end memory networks proposed by Sukhbaatar et al. (2015) for question answering is taken for our experiments. We evaluate the performance of memory networks and the proposed modifications using

the mentioned 4 synthetic datasets. We also check for the effectiveness of attention mechanism in memory networks to aid coreference resolution, through attention mechanism accuracy. We get close to 100% accuracy on all the synthetic datasets, with the best performing modifications on the memory networks. Also memory networks gives a huge accuracy improvement over one of the state-of-the-art coreference resolution system; Cort (Martschat et al., 2015a).

1.8 Thesis Organization

We now discuss how the rest of the thesis is organized. The broad theme of the work and how its organization into sections in each chapter is described.

Chapter 2 discusses the literature survey of the existing or related work of the MIR tasks in this thesis. With regard to motif identification, we highlight several notable existing work for Western music on symbolic representation as well as audio. The novelty of the raga similarity identification make it devoid of any existing work. We discuss the existing work on raga recognition which is the closest work to this. Since there are many aspects to the problem of coreference resolution, the discussion on existing work is more detailed covering different modeling paradigms, mention detection, rule based approaches and statistical approaches.

Chapters 3 and 4 discuss melodic motif detection from Hindustani audio performances. Chapter 3 describes our approach for *mukhda* detection. After introducing the problem, Section 3.1 discuss the dataset and the characteristics of *mukhda*. Section 3.2 discusses extraction of predominant vocal pitch along with the discussion on the main idea including candidate selection and similarity modeling. Section 3.3 give details on the experiments and, describe the results and observations.

Chapter 4 looks into *pakad* detection starting with an introduction to the problem. Section 4.1 describes the dataset, the details of the *pakads* in the chosen raga and on the annotation of ground truth. Section 4.2.1 discusses the pre-processing done on the audio including predominant vocal pitch detection and refinement of segment boundaries. A detailed account on the phrase level characteristics of a *pakad* is given in Section 4.3. Section 4.4 describes the similarity computation steps involving constraint learning for DTW. Experiments and results are discussed in Section 4.5.

Chapter 5 deals with detection of raga similarity from *bandish* notation. Section 5.1 discusses the motivation and the central idea. Section 5.2 describes the proposed neural network

architecture for learning note-embeddings and Section 5.3 details on the note-embeddings and on computation of similarities from the note-embeddings. Section 5.4 discusses the baseline approaches with N-gram, pitch-class distribution and uni-directional LSTM. Section 5.5 details on the dataset, followed by the pre-processing done on the data in Section 5.6. Experiments are discussed in Section 5.7 and the results comparing our approach with the baselines in Section 5.8.

Chapter 6 discusses detection of raga similarity from textual discussions. Section 6.1 details about distributional semantics followed by Mikolov's approach and our proposed modification to Mikolov's approach in Section 6.2. Section 6.3 discusses the dataset. Section 6.4 details the experiments and the evaluation methods including quantitative and qualitative evaluation. Results of both the evaluations are discussed in Section 6.5 and the Section 6.6 describes the results on different datasets to show the applicability of this approach for any specific-domain dataset.

The remaining chapters discuss on information extraction from natural language text content; Chapter 7 gives an introduction to need of information extraction from text for Hindustani MIR and coreference resolution. Section 7.1 gives details of the dataset comprising forum posts from Rasikas.org. Section 7.2 describes our approach of coreference resolution following a supervised mention-pair model. Section 7.3 gives an account of the features used in our approach including the modifications done and the novel features. Section 7.4 discusses our approach for mention detection using regular expressions. Bayesian network based approach for mention-pair classification for a small dataset is discussed in Section 7.5. Section 7.6 discusses the hybrid approach combining machine learning based classifier with a rule-based sieve. The extensive experiment and results Section 7.7 discusses the result of different classification approaches discussed along with evaluation of features. This section also discusses the accuracy of our mention detection approach and compares the results of our system with one of the state-of-the-art systems.

Chapter 8 discusses our proposed idea to improve Best-first clustering. Section 8.1 introduces Best-first clustering with an example. Section 8.2 discusses the motivation and intuition behind the proposed method. This section also presents a formalized algorithm of the proposed modification. Section 8.3 presents the results comparing the proposed modification and its variants with the baseline Best-first clustering.

Chapter 9 discusses semantic class identification for the entities in the Rasikas.org dataset.

This chapter starts with introducing the requirement of semantic class feature and need for a different approach considering the nature and domain of the text content. Section 9.1 looks into semantics for coreference through the existing works discussing different approaches for integrating semantic information for coreference resolution. This section also discusses the motivation and the essence of our approach for semantic class identification. Section 9.2 discusses the semantic classes in Indian classical domain apart from the standard semantic classes in generic domain. Section 9.3 discusses our web based approaches including the baseline approach using Freebase and the proposed approaches using content from the online search results. Section 9.4 compares the results of the discussed approaches and the baseline based on the accuracy of identification. Confusion matrices give detailed information of the class-wise confusions and accuracy. Section 9.5 shows the results on improvement in coreference resolution accuracy when semantic class is integrated as a feature.

Chapter 10 discusses our investigation into novel approaches to bridge the gap between the existing approaches and human process of coreference resolution. Eye-tracking based approach and the approach using memory networks are discussed in two different sections. Section 10.1 starts with the motivation for utilizing eye-tracking for coreference resolution and describes relevant existing work using eye-tracking and the elements of eye-tracking study. This is followed by details on how eye-tracking experiments are conducted to build eye-movement database in Section 10.1.1. Section 10.1.3 describes our method to utilize gaze information for mention pair pruning in automatic coreference resolution. Experiments and results in Section 10.1.4 details the experiments and report the results with different mention-pair classifiers and with one of the state-of-the-art approaches after integrating gaze information based pruning. Section 10.2 begins with discussion on the incapabilities of existing coreference resolution systems in comprehending a discourse and motivates on how the capabilities of memory networks can help in mitigating this. Section 10.2.1 describes the architecture of end-to-end memory networks which we deploy for our experiments. Section 10.2.2 describes the modeling of coreference resolution as a question answering problem and the modifications tried on the memory networks. Section 10.2.3 describes the experiment setup and the synthetic dataset. Section 10.2.4 discusses the results with end-to-end memory networks and the modifications tried. This section also compares memory networks with one of the state-of-the-art systems, Cort and gives a detailed analysis of this comparison.

Chapter 2

Literature Survey

In this thesis, we research on a few different tasks which are identified relevant to developing a recommendation system for Hindustani music. This chapter attempts to set up the stage for explaining our contributions. Here we discuss the existing work related to the tasks investigated upon, *viz.* melodic motif detection, raga similarity detection and coreference resolution. The approaches to motif detection in Western music from symbolic as well as audio input are discussed. Since there is no existing work in raga similarity detection, we study the approaches in related areas including raga recognition in Indian classical music and genre similarity detection in Western music. Coreference resolution being one of the well investigated problems in natural language processing, we try to cover the major aspects of the problem and the existing approaches to the problem. This makes the section on this task elaborated compared to the survey of the other two tasks.

2.1 MIR from Audio: Motif Detection

2.1.1 Symbolic Representation

Symbolic score is a choice for quite a lot researches in Western music MIR, because it is considered parallel to the corresponding performance. Many of the motif detection researches take input as symbolic notation. Juhász (2007) discusses an approach for automatically identifying the motifs from Hungarian fold music using self-organizing map taking dynamic time warping (DTW) with adaptive weights as the distance measure. Cambouropoulos (2006) investigates how musical parallelism can be exploited for melodic segmentation. In this method, all the

discovered significant patterns give clue on the likely segmentation points for a pattern. Geometric methods consider a melody as a shape in n-dimensional space where repeated patterns have identical shapes. Meredith's Structure Induction Algorithms (SIA) Meredith et al. (2002) follow a geometric approach and Collins et al. (2010) describes improvements to it. Through shapes, imitations and gestalts and similarity relation between any two shapes, Buteau and Mazzola (2000) creates a motivic topology of a melodic score. Szeto and Wong (2006) presents a graph-theoretic approach modeling a piece of music as a graph. The method described here to search for a pattern in a piece is equivalent to searching a subgraph referred to as maximal matched CRP subgraph.

The motif detection method proposed by Nieto and Farbood (2012) extracts potential motifs through a set of rules followed by its clustering. The distance between two motifs have to be within the defined threshold for them to be clustered together. The idea proposed in Rolland (1999) finds all the equipollent passage couples followed by a categorization phase which extracts actual patterns from a similarity graph. The approach by Meek and Birmingham (2001) identifies the patterns relevant to the theme of the song from the large number of extracted patterns using perceptually significant characteristics.

2.1.2 Audio

It is more challenging to detect patterns from audio; the lowest level of music representation. Considering the effectiveness of the state-of-the-art techniques for transcription, many of the approaches rely on finding patterns after conversion of the audio to some intermediate representation which includes predominant vocal/instrument pitch. Dannenberg and Hu (2003) discusses extraction of pitch information from monophonic and polyphonic audio and representation-specific algorithms to identify patterns. Spectral based analysis discussed in this paper depends on chroma features for audio instances on which obtaining higher level pitch information is difficult. An adaptive method is proposed by Lu et al. (2004) to identify patterns from self similarity matrix based on constant Q transform features extracted from audio. From the obtained repetitions, the structure of music is analyzed with certain heuristic rules. Aucouturier and Sandler (2002) uses timbre information for segmentation, giving a string representation with which repeating patterns are identified using methods inspired by image processing techniques; kernel convolution and Hough Transform. Rafii and Pardo (2013) identifies repeating patterns in the accompaniment in-spite of the varying vocal overlaid. The foreground vocals is separated

and repeating patterns are identified using time-frequency masking. While this method is restricted to identification of short excerpts and does not allow variations, Liutkus et al. (2012) extended this method to overcome these limitations. Peeters (2003) investigated deriving musical structures from audio using static features (MFCC, chromagram) and the proposed dynamic features based on Short Time Fourier Transform. With the aim to generate audio and visual summaries, they study the state and sequence representation for music structure detection using the mentioned features. The data-driven approach by Weiss and Bello (2011) introduces a probabilistic variant of convolutive non-negative matrix factorization for extraction and localization of repeating patterns in an audio.

Logan and Chu (2000) discusses extraction of similar patterns for the purpose of key phrase identification with the help of Mel-spectral features. A clustering and an HMM based approaches are experimented for labeling the frames. Cooper and Foote (2002) finds the representative excerpt from a music audio using self-similarity analysis with MFCC and spectral features. The subsequence having maximum similarity to the whole audio is picked as the representative excerpt.

2.2 Raga Similarity Detection

To the best of our knowledge no such attempts to identify raga similarity have been made so far. Raga recognition is the closest research problem in Indian classical music, which identifies the raga of a given performance. In Western music, analyzing similarities between genre is analogous to our problem, and a few approaches has been introduced.

In raga recognition, Bhattacharjee and Srinivasan (2011) discusses raga identification of Hindustani classical audio performances through a transition probability based approach. Here they also discuss about validating the raga identification method through identifying known raga relationship between 10 ragas considered for this work. A good number of research works have been carried out pertaining to raga identification in Hindustani music using note intonation. Belle et al. (2009) discusses an approach extracting features from swara intonation to be used in a statistical classification framework. This work uses a dataset with Hindustani ragas from the same *thaat* (sharing same set of notes) to validate the relevance of intonation variabilities of the same swaras in different ragas to raga recognition. Dighe et al. (2013) investigates features consisting of chromagram patterns along with mel-cepstrum coefficients and timbre features,

with a dataset containing performances from ragas belonging to different *thaats*. The same paper also discusses swara histogram based approach giving comparable, but lower accuracies w.r.t the former. Tansen (Pandey et al., 2003) system utilizes an automatic note transcriptor to extract swaras from an audio performance to perform raga recognition. Hidden Markov Model enhanced with a string matching algorithm is employed in this system. Tansen is experimented with a dataset having performances only from 2 ragas, but from the same *thaat*.

There are also a good number of approaches for raga recognition in Carnatic music. The approach by Sridhar et al. (2011) uses Latent Dirichlet Allocation (LDA) for raga recognition, which is widely used for document classification. Here, a swara corresponds to a word and the performance to a document. The dataset comprises of performances from parent ragas as well as child raga and the swara sequences are extracted with an automatic transcription method. Bellur et al. (2012a) studies the importance of motivic analysis for raga recognition. The proposed HMM based classifier is designed to classify the raga of a motif, to which the motif belongs to as a characteristic phrase. Shetty et al. (2012) discusses an approach for clustering raga performances through analyzing the jump sequences of swaras in the performances. Padmasundari and Murthy (2017) presents an approach using Locality Sensitive Hashing (LSH) that can hash similar items to the same bucket. This paper claims that, this approach better models human perception of ragas. In the approach by Ranjani et al. (2011), the authors have proposed a semi-continuous stochastic mixture model to identify the tonic frequency. The approach incorporates the template structure (Sa-Pa-Sa) to identify the tonic and in-turn the statistics of other notes. From this, the raga recognition approach for *melakarta* ragas (those which contain all 7-notes) follows by determining the most likely set of notes.

In western music the closest counterpart to the task of raga similarity identification is identification of similarity between music genres. Perhaps due to the limited number of genres and the inter-genre similarities are known, this problem has not been a primary concern to the main stream MIR research. Still, we can see a few approaches for analyzing similarity between genres which are developed primarily to help genre classification and related problems. With the aim to reduce inter-genre confusion in genre classification, there are a few attempts to model inter-genre similarity (Bagci and Erzin, 2009; Erzin et al., 2007). Inter-genre similarity modeling relies on timbral texture features. On finding that music stretching resistance has impact on intra-genre cohesiveness and inter-genre discrepancies, Chen and Wang (2017) investigates on identifying genre similarities from music stretching resistance of different genres.

2.3 Coreference Resolution

coreference resolution has been researched for many years, where the initial approaches were mostly knowledge-driven rule based approaches. These approaches continued to be dominate till the advent of data-driven approaches. Along with discussing the existing approaches in rule-based and data-driven paradigms, we will also be discussing the linguistic aspects of coreference resolution. Discussions will also be extended to different modeling paradigms for the task of coreference resolution and some existing work in mention detection.

2.3.1 Linguistic and Other Considerations

Linguistic factors such as syntactic constraints, semantic cues like gender, world knowledge and knowledge of textual structure are of primary importance to resolution of coreference (Crawley et al., 1990). Most of the widely discussed features which are found effective for the task, are motivated by these factors. Interpretation of noun phrases in many cases depend on the linguistic context, considering the discourse situation (Bean and Riloff, 2004). The domain of interpretation is a key to finding relation between mentions.

Syntactic structure and syntactic preferences play a major role. Majority of the pronouns and their antecedents occur in the subject position of the sentence. Kertz et al. (2006) discusses parallel function preference stating that an anaphora and its antecedent tend to have the same grammatical role. An anaphoric mention with a subject grammatical role is likely to have an antecedent with subject role. This paper also discusses subject preference, where an antecedent is likely to be having subject role, except for the cases when there is a gender or number mismatch. Government and binding theory are key in constraining pronoun resolution especially for resolution between mentions in the same sentence (Chomsky, 1982). Hobbs (1978) discusses the parse tree based constraints c-commands and s-commands. Barker and Pullum (1990) explores tree based constraints like the check for whether the anaphora and antecedent are sister nodes, or they are the subject/object of the same verb.

When different candidates compete to be the antecedent for an anaphoric mention satisfying most of the criteria, the candidate which has the most focus is given preference. This is termed as centering. In the sentence *Hari kept the cup on the table and broke it*, *it* refers to the cup which is the center of attention (Mitkov, 1999). Salience is also an important factor in anaphora resolution. Mostly, the antecedent for an anaphora appears in the same sentence or in

the previous sentence (Hobbs, 1978).

Resolving coreferent relations in some contexts depend on common sense knowledge. See the difference in association of *they* in the following sentences.
The city council refused the women a permit because they feared violence.
The city council refused the women a permit because they advocated violence.
Here the verb determines the association of *they* (Webber and Joshi, 1998). Semantic parallelism constraints the coreferent mentions to have the same semantic role and this has precedence over syntactic parallelism. syntactic and semantic parallelism act as filters to filter out unsuitable candidates (Kertz et al., 2006).

Morphological constraints including gender, number and person constraints are crucial to resolution (Ehrlich and Rayner, 1983). For example, in the sentence *John met Mary to discuss his problem*, *his* is referring with *John* and in the sentence *John met Mary to discuss her problem*, *her* is associated with *Mary*. Here the resolution is governed by the gender of the mentions.

2.3.2 Mention detection

Mention detection finds the mention boundaries in the corpus, identifying named entities, nominal and pronominal references. Mention detection is a crucial step in coreference resolution, since its accuracy directly contributes to the accuracy of the coreference resolution system. Zhekova and Kübler (2010) describes a method to identify mentions from named entities, pronouns and noun phrases based on certain constraints. There has been a few supervised approaches for mention detection. Nicolae and Nicolae (2006) uses maximum entropy classification to classify a noun phrase as mention or not, whereas Sikdar et al. (2015) uses a CRF based approach for identifying whether each word is part of a mention. Both these approaches uses simple lexical and syntactic features.

Mention detection is sometimes task dependent in order to bring in the peculiarities pertaining to that particular task. Pradhan et al. (2007b) analyzes mention extraction from parse trees by extracting noun phrase, pronoun, pre-modifier and verb. The difference in mention detection accuracy when using gold standard trees and automatic parse trees is discussed in Yang et al. (2011). Björkelund and Farkas (2012) depends on syntactic trees as well as named entities. Chen and Ng (2012) implements an extraction step, extracting named entities and other mentions through language specific heuristics. This is followed by a pruning step consisting language-specific heuristic pruning and language independent learning based pruning. Durrett

and Klein (2014) discusses a joint model for coreference resolution, named entity identification and entity linking, motivated by the interdependencies between these tasks.

2.3.3 Rule Based Approaches

Most of the initial approaches in coreference resolution were highly linguistic oriented till the introduction of data-driven approaches. Winograd (1972) proposed a coreference resolution system as part of an automated English understanding system, considering all preceding noun phrase candidates for probable antecedent and rate them based on their syntactic position. Hobbs (1978) discusses one of the earliest syntactic approaches; Hobbs algorithm. For a pronominal mention, this algorithm utilizes constituency parse tree to identify the antecedent. This method also incorporates syntactic constraints and semantic considerations through rules.Rich and LuperFoy (1988) scores the probable antecedents after evaluating each antecedent with a set of defined constraint sources. The final score of an antecedent candidate is a function of score given by each constraint source and the confidence associated with the constraint score.

CogNIAC system (Baldwin, 1997) resolves a subset of anaphoras without requiring much of world knowledge and complex linguistic processing. The rules are constructed in such a way that it reproduces the human way of pronoun resolution. These rules are formed taking into consideration the nature of anaphora and antecedent and the nature of text containing the mentions. For example, if anaphora is reflexive then pick the nearest antecedent from the current sentence . Lappin and Leass (1994) explains a rule based approach, applying different rules after generating syntax representation of text using McCords slot grammar parser (McCord, 1990). Intra-sentential filter is responsible for identifying non-coreferential mentions based on syntactic features including filtering out of mentions with incompatible agreement features (gender, number etc.) and mentions argument of the same phrasal head. Other filters include morphological filter to filter out incompatible pronoun, NP coreference, filter to identify pleonastic pronoun. Kennedy and Boguraev (1996) presents modifications to this approach using POS tags and grammatical function of lexical items instead of syntactic parse information. Mitkov (1998) presents a knowledge poor approach employing information available from a part-of-speech tagger and with certain simple noun phrase rules. Here, for an anaphoric reference, search for the antecedent is restricted to the current sentence and 2 previous sentences. The candidate noun phrases in these sentences which agree on gender and number with the

anaphora, are scored with antecedent indicators.

Even after the introduction of data driven approaches in coreference resolution, there were a few rule-based systems exhibiting matching performance compared to the state-of-the-art systems. Among these, the prominent one is the Stanford coreference resolution system (Raghunathan et al., 2010) based on a multi-sieve rule-based approach. This applies deterministic coreference models at different phases in the descending order of their strength in deciding coreference. The initial passes resolve exact match, appositives, relative pronoun etc. The last pass is dedicated to resolve pronouns.

2.3.4 Data-driven Approaches

Features for Coreference

Features provide the essential clues for checking coreferent relation between mentions in any machine learning based approach. These features are strongly motivated by the linguistic clues for coreference. Mostly the features are computed taking two mentions at a time; except for the approaches where the belongingness of a mention to a cluster of mentions is evaluated.

As per Ng and Cardie (2002a), features are generally classified as lexical, grammatical (NP type, NP property/relationship, syntactic pattern), semantic and positional. Uryupina (2006) also illustrates a similar classification of features based on lexicographical similarity, syntactic knowledge, semantic compatibility and discourse & salience. Semantically compatible features attempts to incorporate semantic knowledge, where as discourse and salience features checks for distance, salience based on linear order, hierarchy of semantic roles and centering parameters. The perspective of categorization of features in Recasens and Hovy (2009) is different. Classical features captures the basic characteristics of two mentions to be coreferent. This includes check for pronoun, proper noun, comparison with Wordnet for similarity analysis. Language specific features take care of the language specific aspects of coreferent relation, like in Spanish where elliptical subjects and grammatical gender are important for determining coreference. Corpus specific features are dependent on the definition of coreference w.r.t the corpus. For instance, AnCora corpus (Recasens and Martí, 2010) considers identity relation different from apposition and predication relations. Novel features proposed include features to capture the joint behaviour including whether both the mentions are subjects in the sentences. It includes features to check counter match, which prevents mentions having different numerals to

corefer (ex. '134 million euros' and '100 million euros' does not corefer). Kobdani and Schütze (2010) categorizes the features into two types viz., atomic features and link features. Atomic features include sentence number of word, grammatical gender, POS tag, pronoun person capturing the specificities of the word/phrase comprising the mention. Link features capture the relation between the pair of mentions, including the features that checks for string similarity between the mentions and whether the mentions are in the same sentence.

Lexical features are of high significance to coreference resolution (Recasens and Hovy, 2009). The string similarity can be a sole reason for the mentions to be coreferent. In this category, the features to check if two mentions have high string similarity, if one mention is the alias (eg. abbreviation) of the other are widely used (Ng and Cardie, 2002b; Rahman and Ng, 2009; Soon et al., 2001; Zheng et al., 2011; Recasens and Hovy, 2009; Ponzetto and Strube, 2006). Head match which checks for whether the mentions share the same head is also identified as an essential feature. Certain modifications to string matching feature are discussed by Rahman and Ng (2009) proposing checks for whether both the mentions are pronominal or proper noun, along with the check for string similarity. Similarly, Ng and Cardie (2002b) discusses the same set of features along with substring matching. Instead of strict string matching, edit-distance variations are employed by many systems considering the nature of the text (Pradhan et al., 2007b; Zheng et al., 2011; Strube et al., 2002; Stamborg et al., 2012).

Appositive feature helps to identify coreferent mentions which are coreferent due to the attribution of one mention on the other (Recasens and Hovy, 2009; Soon et al., 2001; Rahman and Ng, 2009). In the sentence "Ram Nath Kovind, the president of India made an important decision.", there is an appositive relation between the mentions '*Ram Nath Kovind*' and '*the president of India*'. Distance based features belongs to the commonly used set of features, including sentence distance that computes distance between the mention in terms of number of sentences in between them, word distance computing the number of words, and mention distance computing the number of other mentions (Recasens and Hovy, 2009; Soon et al., 2001; Rahman and Ng, 2011b; Ding and Liu, 2010). Features to check for the type of the mention noun phrase (NP) are extensively experimented. This includes checks for definite noun phrase (NP starting with 'the'), indefinite noun phrase (NP starting with 'a', 'an') and demonstrative noun phrase (NP starting with 'this', 'that', 'these', 'those'). Likewise, the features to check for whether a noun phrase is pronoun or proper noun are inevitable to any coreference resolution system (Soon et al., 2001; Rahman and Ng, 2009; Recasens and Hovy, 2009). Gender and

number agreement features have also been part of many coreference resolution systems (Soon et al., 2001; Recasens and Hovy, 2009; Ng and Cardie, 2002b), except for certain domain-specific systems where it is difficult to extract these features.

Grammatical role features helps to capture the relative relevance of mentions. It is very likely that a pronoun will be coreferential with an antecedent having the same grammatical role (Grober et al., 1978). Crawley et al. (1990) states that a pronoun has more chances of being coreferential with the subject of the preceding clause. Information obtained from grammatical role of the mentions, syntactic paralellism or collocation patterns by syntactic features are of high significance to coreference resolution (Weissenbacher and Sasaki, 2013). Rahman and Ng (2011c) discusses extraction of path based features by analyzing the nodes in between the mention and potential antecedent in the syntactic parse tree. To analyze relation between mention lying in different sentences, parse trees of different sentences are combined through a pseudo-root node. Tree based features are made use in a tree kernel, forming a composite kernel with flat features and tree based features training an SVM classifier. In the work by Lassalle and Denis (2013) linguistic concepts such as governing category, apposition and dependency relationships are extracted from syntactic parse trees. Along with that features inspired by binding theory which helps to find antecedent of a reflexive or reciprocal pronoun within the sentence is discussed.

Coreference resolution is one of the problems in which pragmatics and world knowledge play a key role (Recasens and Hovy, 2009). Bringing forth semantic knowledge and world knowledge is one of the biggest challenges in coreference resolution. In this category, semantic class agreement between the mentions turned out to be one of the most discussed feature. Semantic classes encompass *person, organization, location, date, time, money, percent*, and *object* (Recasens and Hovy, 2009; Soon et al., 2001; Rahman and Ng, 2009). Animacy feature (Recasens and Hovy, 2009; Rahman and Ng, 2009), a generalized variant of semantic class is simple to be extracted automatically . To automatically determine the semantic class, a named entity recognizer or a Wordnet (Miller, 1995) based recognizer is employed (Rahman and Ng, 2009; Recasens, 2010; Nicolae and Nicolae, 2006). There also exist certain other Wordnet based features viz., Wordnet distance (Recasens, 2010; Recasens and Hovy, 2009), similarity score (Raghavan et al., 2012; Uryupina, 2006; Ponzetto and Strube, 2006), synonymhypernym relation (Hendrickx and Hoste, 2009; Ng, 2007a), word-sense agreement (Zhou and Kong, 2009) etc. Wikipedia is another important source of external knowledge which has been widely used.

Ponzetto and Strube (2006) discusses an extensive list of features derived from Wikipedia for a pair of mentions. Among the two sets of features discussed, one set derives the relatedness between the mentions from the Wikipedia page content, and the other set computes the relatedness from the closeness in the Wikipedia categories of the mentions. A few other approaches extract information from Wikipedia to identify the linkage between the mentions (Durrett and Klein, 2014; Cheng and Voigt, 2015). Apart from these sources of world knowledge, there are some attempts to utilize a few other knowledge sources. Rahman and Ng (2011a) proposed the use of Yago (Suchanek et al., 2007) and Framenet (Baker et al., 1998), where Yago is used for extracting is-a and is-also-known-as relations, and from Framenet information on co-occurrence of predicates having similar semantic role labels are extracted. Bansal and Klein (2012) captures features for coreference semantics from web resources through co-occurrence and contextual clues.

Apart from these features, different coreference resolution systems have proved the usefulness of various other features. Stamborg et al. (2012) discusses incorporating linguistic phenomena and discourse properties to the features. They discuss some novel features including discourse and type of document . For coreference resolution in certain languages (eg. Spanish), feature to check if a mention is an elliptical pronoun is crucial (Recasens and Hovy, 2009). Rösiger and Riester (2015) discusses prosodic features for resolving coreference in spoken text.

Modeling Coreference Resolution

Since the decision of coreference involves many mentions in a text, there are different ways the problem can be modeled. Mostly in all these methods, at the root level the comparison is between the two mentions at a time. There were different attempts to model the problem of coreference resolution. Some existing approaches experiment with different coreference models to demonstrate the impact of their contribution.

Mention-Pair Model: Mention-pair model has a classification step followed by clustering. Classification take into consideration two mentions at a time, classifying them as coreferent or not. For an anaphoric mention m_k, the classification step checks if a candidate antecedent m_k is coreferent (Rahman and Ng, 2009). Features are computed for each mention pair. In a supervised approach a training instance is created with a mention and its closest antecedent, and for the same anaphoric mention negative instances are created by pairing with mentions occurring before it and after its closest antecedent (Soon et al., 2001). During testing, the clustering step

following the classification, identifies the best antecedent for an anaphoric mention. Clustering picks the best antecedent from the candidate antecedents which are identified coreferent with the anaphoric mention after classification. This forms distinct coreferent chains in a document.

Entity-Mention Model: When classification confines to a mention pair in the mention-pair model, entity-mention model compares with previously identified partial clusters. The classifier determines if a mention belongs to one among the partial clusters occurring before this mention. Each training instance contains a mention and a cluster and the computed features include cluster level features and features pertaining to the mention under consideration (Luo et al., 2004; Yang et al., 2008).

Mention-Ranking Model: In this model the candidate antecedents of an anaphoric mention m_k are ranked. Instead of a classifier as in mention-pair model, a ranker learning algorithm is chosen for this. A training instance $i(m_j, m_k)$ represents m_k and a preceding mention m_j. S_k represents the set of training instances ($(m_j, m_k), m_j \in candidate_antecedents$) for an anaphoric mention m_k. Each instance in S_k is given a class value which is the rank assigned to the candidate antecedent in the instance. Rank of 2 is given if m_j is the closest antecedent of m_k and 1 otherwise. After training, the mention-ranking model is applied to the test instances (Denis and Baldridge, 2008; Rahman and Ng, 2009).

Cluster-Ranking Model: Cluster ranking model is a hybrid of mention-ranking model and entity-mention model, which ranks preceding clusters for an anaphoric mention m_k. A training instance is created from a discourse-old mention and its preceding clusters, and cluster level features are extracted. For a training instance $i(c_j, m_k)$, class value of 2 is given if m_k belongs to the cluster c_j, and 1 otherwise. Since the training instances do not contain discourse new mentions, this method is not equipped to find discourse-new mentions. Since the training instances involves only mentions having antecedent, this model fails in identifying discourse-new mentions in a document. To get around this, the ranker has to be trained with instances generated from discourse-old and discourse-new mentions (Rahman and Ng, 2011b, 2009).

Supervised Approaches

Supervised approaches gained popularity by mid-1990 in resolving coreference. Based on the aforementioned ways to model the task of coreference resolution, the machine learning approaches can be broadly classified into two; one is a 2 step approach with a binary classification followed by clustering and the second is a ranking approach (Zheng et al., 2011).

One of the earlier statistical approaches is by Dagan and Itai (1990), where word co-occurrences are taken into account to disambiguate pronouns, but restricted to the pronoun *it*. For an *it* coming as subject of a verb, the mention among the candidate antecedents having maximum co-occurrences with the same verb as subject is selected as the antecedent. In another different attempt by Ge et al. (1998) proposed a statistical framework for resolution of third person pronouns which learns a probabilistic model using Penn Wall Street journal Treebank (Riezler et al., 2002). For a candidate antecedent to an anaphoric mention, this computes probability values for certain factors (eg. distance, co-occurrence patterns etc.) based on the probability values computed over the training data. These probabilities are multiplied to compute the probability associated with a candidate antecedent.

Introducing mention-pair model, one of the widely used modeling paradigm for coreference resolution, Aone and Bennett (1995) introduced a coreference resolution system for Japanese coreference resolution. They experimented C4.5 decision tree classifier for mention pair classification. Different variants of their approach are evaluated against their own previously designed solver based on manually selected knowledge sources. During the same time, McCarthy and Lehnert (1995) and Connolly et al. (1997) came up with a machine learning based approach for English. Following the same modeling paradigm of Aone and Bennett (1995), Soon et al. (2001) built a machine learning based coreference resolution system focusing more on the design of features. They employed C5 decision tree algorithm for mention pair classification, and the system gives matching performance with the then existing rule-based systems on MUC-6 and MUC-7 datasets. Extending this work, Ng and Cardie (2002b) introduced a deeper set of features for coreference. Ng and Cardie (2002a) modified this approach by determining the anaphoricity of a noun phrase as a pre-processing step. Bergsma et al. (2008) determines the anaphoricity of a noun phrase through a method based on context distribution, Ram and Devi (2012) discusses a CRF based approach for determining anaphoricity, and Ng (2009) proposes a graph-cut based anaphoricity determination algorithm. Uryupina (2006) experimented with different classifiers extending the feature set from the conventional set of features with more linguistically motivated features. There has been several attempts to improve the discussed methods through utilization of semantic knowledge from diverse sources. Along with introducing Ontonotes; the present widely used dataset for coreference resolution, Pradhan et al. (2007b) introduced a baseline model with classifier as Support Vector Machine.

Several researches considered on finding alternatives to mention-pair model, leading to

introduction of entity-mention model (Luo et al., 2004), mention-ranking model (Denis and Baldridge, 2008) and cluster-ranking model (Rahman and Ng, 2011b). Experiments with ACE 2005 shows that entity-mention model do not give better performance over mention-pair model with system mentions, but gives a small performance improvement with gold mentions. Mention-ranking and cluster-ranking approaches give significant improvement over mention-pair model in these experiments (Rahman and Ng, 2009).

The diverse knowledge sources required for resolving coreferent mentions motivated introduction of many hybrid approaches combining a rule-based and machine learning approaches. Hendrickx et al. (2007) combines a rule-based filtering mechanism having linguistically motivated positive and negative filters with machine learning for a Dutch dataset. The hybrid approach by Chen and Ng (2012) incorporates lexical information through machine learning technique which is combined with a multi-sieve rule-based approach. The CORUDIS system (COreference RUles with DIsambiguation Statistics) (Hartrumpf, 2001) combines a machine learning based method with syntactico-semantic rules. Sobha et al. (2011) uses salience based measures to resolve pronominal mentions and a CRF based method to resolve the non-pronominal mentions. The hybrid approach investigated by D'Souza and Ng (2012) is for coreference resolution in biomedical literature.

There are approaches which model coreference resolution as a graph problem where the nodes represent mentions and the edges represent the relation between the mentions. In most of these approaches, the classification confidence between the mentions is taken as the weights and a graph partitioning algorithm finds the coreferent chains. Nicolae and Nicolae (2006) devised a graph partitioning based method inspired by min-cut graph partitioning algorithm. This approach starts with assuming that all mentions belong to the same entity and divides until the stopping criterion is satisfied. The coreference confidence values which are represented as the weights in the graph are obtained from a maximum entropy model trained on the training documents. Martschat et al. (2012) discusses a greedy clustering, where the edge weights are computed using simple descriptive statistics on training dataset. In their research on event coreference resolution, Chen et al. (2013) present a modified random walk model to perform graph partitioning, where mention pair classification is done by SVM with syntactic structures incorporated through a convolution tree kernel.

Probabilistic models based methods helped in modeling the problem differently. As opposed to other approaches, the pair-wise decisions are not made independently to avoid incon-

sistencies. McCallum and Wellner (2005) proposed a probabilistic model defining a set entities (E) to which the mentions refer to. The set of mentions (X) and their entity assignments (Y) are taken as random variables and there is a random variable associated with the collection of all attributes over all the entities (A). A conditional probabilistic model is trained to maximize P(Y,A|X). They have tried different variants of this modeling. In a later work, Culotta et al. (2007) followed similar approach enabling features over a set of noun phrases. They show how first-order logic can be utilized for this modeling to compute features for a set of mentions. Wick et al. (2012) discusses a discriminative hierarchical probabilistic model, where an entity is partitioned as a tree of latent sub-entities with the mentions as leaves and its associated attributes. A similar approach is used for cross-document coreference (Singh et al., 2011).

Markov logic network (MLN) combines first-order logic with probabilistic graphical model, associating a weight with each first-order logic formula (Richardson and Domingos, 2006). Singla and Domingos (2006) use MLN for entity resolution with first-order logic rules, mainly for word based similarities. To mitigate the inconsistencies while resolving pronoun and noun phrases through a single process or two separate processes, Huang et al. (2009a) proposed an MLN based solution to handle pronoun and noun phrases together through introduction of appropriate first-order logic rules. Song et al. (2012) proposed an MLN based method for joint learning, combining pairwise classification and mention clustering. In this work, the first-order logic rules include separate formulas for pairwise classification and mention clustering.

Recently, there were a few studies on applying deep learning for coreference resolution. The approach by Godbole et al. (2015) investigates how neural embeddings of words and sentences can be utilized along with a set of baseline features. Clark and Manning (2016) compare approaches based on two different clustering methods; mention-ranking and cluster-ranking, giving the then best results with CoNLL 2012 shared task dataset. These approaches discuss neural network models to learn representation for a mention pair and for a cluster pair, to be used with the respective approaches. The most recent work proposing an end-to-end neural coreference resolution system reports a significant improvement over the existing systems and gives the best results till-date on CoNLL 2012 shared task dataset (Lee et al., 2017). The first step in this approach learns a representation for the possible spans in a document with the help of a bi-directional LSTM based neural network and the neural network designed for the second step identifies the best antecedent for an anaphoric mention, scoring the anaphoric mention and candidate antecedents pairs.

2.4 Summary

This chapter presents a literature survey of three different tasks we have explored for extracting music related information from audio as well as text content. As mentioned, in our research, the task of motif detects extracts information from audio and coreference resolution extract information from text content. Different approaches are designed for raga similarity detection to extract similarities from music notation data and textual discussions. Our discussion on the existing approaches to motif detection covers detection from both audio and symbolic content. Considering the novelty of the problem of raga similarity detection, the discussion on the existing work confines to the widely researched problem of raga recognition, and detection of similarity between Western music genres. While discussing the existing approaches to coreference resolution, we also discussed other aspects of the problem. We discussed different modeling paradigms in coreference resolution along with the linguistic considerations relevant to the automation of the problem. We discussed approaches to mention detection, a crucial component of any coreference resolution system.

Chapter 3

Mukhda Detection

Melodic motifs form essential building blocks in Indian Classical music. As the *pakads* are important to the identity of a raga, the main title phrase of a *bandish*; *mukhda* is decisive to a *bandish* . A *mukhda* can be found recurring in a *bandish* performance. The occurrence of *mukhda* has association with the rhythmic cycle; *mukhda* is rendered in a way that a particular syllable of *mukhda* is always rendered at the strongly acccented down beat (*sam*) of the rhythmic cycle. In the improvised section of the concert known as the *bol-alap*, the singer elaborates within each rhythmic cycle of the tala using the words of the *bandish* interspersed with solfege and held vowels, purposefully reaching the strongly accented first beat (the *sam*) of the next rhythmic cycle on a fixed syllable of the signature. Even in the segments where the singer makes his own improvisations like *bol-alap* and *taan*, the singer tries to deliver *mukhda* strictly observing its rhythmic association. Here we discuss a method to automatically identify *mukhda* instances in a *bandish* performance. The relation of *mukhda* to the rhythmic cycle serves as a strong cue in this task. We focus on the melodic and rhythmic invariances to provide cues for the automatic detection of all occurrences of the *mukhda*, provided one reference instance, across the audio recordings of *bandish* of prominent artistes.

In this work, we consider the segmentation of audio signals with the available cues and detecting the *mukhda* instances by computing similarity measures on time series of automatically detected pitch values. Given the well-known difficulties with extracting low-level musical attributes such as pitch and onsets from general polyphonic audio recordings, most of the work in motivic analysis for Western music has been restricted to symbolic scores. Considering the characteristics of Hindustani music, for this work, we depend on a predominant pitch detection algorithm (Rao and Rao, 2010) to extract the melodic pitch contour from vocal concert record-

ings. For a particular performance, candidate phrases are spotted from all the identified *sam* occurrences with the help of information given on the *mukhda* duration and its relation with a *sam* instance. For detecting *mukhda* phrases from these candidates, we experiment with different similarity measures which are widely used for time series matching. The proposed motif detection method is evaluated for within concert and across concerts (of the same *bandish*) detection. Considering the importance of a *mukhda* phrase to its *bandish*, detection of *mukhda* has applications to retrieval of *bandishes*.

The main focus of this chapter is on our method for *mukhda* detection. This starts with describing the dataset and the characteristics of *mukhda* in Section 3.1. Section 3.2 discusses the main steps including extraction of predominant vocal pitch, candidate selection and similarity modeling. The proposed method for candidate selection and different similarity measures considered are detailed in this section. Section 3.3 gives details on the experiments and, describe the results and observations.

3.1 Dataset and Evaluation Methods

We selected four full-length CD-quality recorded concerts of well-known Hindustani *khyal* vocalists. In all cases, the accompanying instruments are tanpura (drone), harmonium and tabla. The section of each concert corresponding to *bandish*-based improvisation (*bol-alap*) is extracted for this study. Table 7.1 shows the artiste names and *bandish* titles with other relevant details including CD cover meta-data and the duration of the *bol-alap* section. All the performances use the popular *tintal* rhythm cycle with 16 beats divided equally over 4 sections. The beats are realized by the strokes of the tabla (percussion), with the first beat of each section considered to be stressed in 3 of the 4 sections. All the *mukhda* phrases which may occur around any *sam* (first beat of the cycle) throughout the performance, are manually labeled. This serves as the ground truth (positives) for the motif detection evaluation. The tempo indicated for each piece is an average, with slow fluctuations in cycle length observed throughout the recordings. Of the four recordings in Table 7.1, the first two correspond to the same *bandish* by different artistes. The last recording is by a female vocalist. It was observed that this recording with its slow tempo exhibits the largest variations in the duration of the *mukhda* even after accounting for local variations in cycle length.

For further processing, the audio is converted to 16 kHz mono at 16 bits/sample. Figure

Artiste	Raga	Tala	Bandish	Tempo (bpm)	Dur (min)	#Phrases	
						Positive	Negative
Bhimsen Joshi (BJ)	Marwa	Tintal	Guru Bina Gyan	193	4.58	13	55
Ajoy Chakraborti (AC)	Marwa	Tintal	Guru Bina Gyan	205	9.08	33	295
Bhimsen Joshi (BJ)	Puriya	Tintal	Jana na na na	204	9.36	17	97
Kishori Amonkar (KA)	Deshkar	Tintal	Piya Jaag	43	22.3	44	176

Table 3.1: Description of dataset

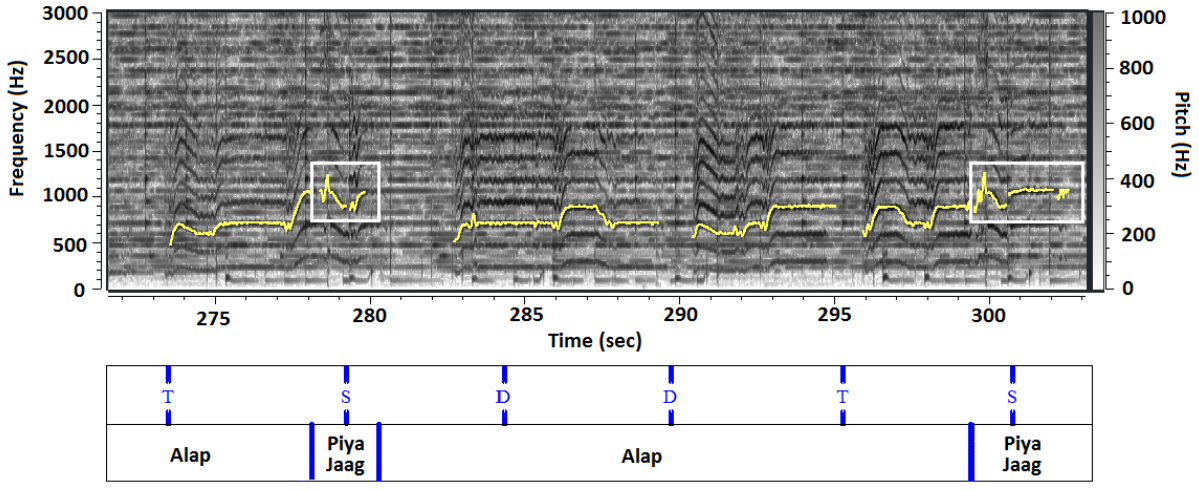


Figure 3.1: Top: spectrogram with superposed vocal pitch and *mukhda* in boxes; bottom: first beat of each subcycle (S= *sam*) with aligned lyrics in vocal regions.

3.1 shows the spectrogram (of the 3 kHz frequency range) for a duration slightly greater than 1 full rhythmic cycle extracted from the *Piya Jaag* recording by Kishori Amonkar. Superimposed on the spectrogram is the detected pitch contour (as obtained by the method presented later in Sec. 4). Beneath the spectrogram is an annotation block depicting the aligned *tala* cycles. The first beat of the cycle is the *sam* (S) corresponds to the *dha* stroke of the tabla. In Figure 3.1, the first beat of each sub-cycle is labeled (*dha* (D) or *tha* (T)). The penultimate sub-cycle before the S is the *khali*, as also evident from the absence of low frequency tabla partials in the spectrogram of this segment. The *mukhda* segments corresponding to the utterance "*Piya Jaag*" are enclosed in boxes. The *mukhda* segments are observed to be melodically similar and also similarly aligned within the *tala* cycles. Note that the song syllable that coincides with the *sam* (S) is sometimes left incomplete by the vocalist.

3.2 Automatic *Mukhda* Detection

The pitch contour depicted in Figure 3.1 can be viewed as a time series in which the desired phrase segments are embedded. As such, finding segments in the overall contour that are similar to a given phrase would involve matching the pitch at every time instant of the given phrase to the pitch at every other time instant throughout the time series (Dannenberg and Hu, 2003). It is of interest to explore methods to reduce the search complexity. In the present context, we can exploit the additional knowledge about the rhythmic relationship. As discussed, the vocalist embeds the *mukhda* phrase in the metrical cycle (*tala*) so that a fixed syllable coincides with the *sam* instant. The metrical space of each cycle is occupied by improvisation culminating with the *mukhda*. Motivated by this, we approach the automatic detection of the *mukhda* phrase from the audio by first identifying a limited set of candidate phrases based on the detected rhythm cycle structure, and then computing a melodic similarity distance between the reference template and each of the candidates.

As in any classification task, it is necessary to design an appropriate data representation and a suitable similarity model for the matching. In this section, we describe the steps involved in *mukhda* identification as in Figure 3.2. Finally, candidate segments with distances from the reference template lower than a threshold are the detected positives.

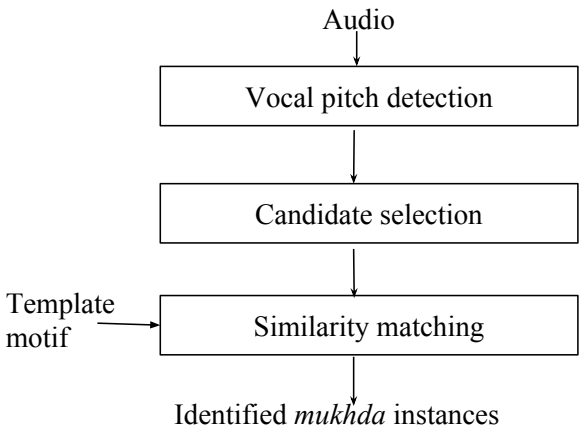


Figure 3.2: Steps in *mukhda* identification

3.2.1 Vocal Pitch Detection

In Hindustani classical vocal music, the accompanying instruments include a drone (tanpura), tabla, and often, harmonium as well. The singing voice is usually dominant and the melody can be extracted from the detected pitch of the predominant source in the polyphonic mix. Melody detection involves identifying the vocal segments and tracking the pitch of the vocalist. We employ a predominant-F0 extraction algorithm designed for robustness in the presence of pitched accompaniment like drone and harmonium (Rao et al., 1999). This method is based on the detection of spectral harmonics helping to identify multiple pitch candidates in each 10 ms interval of the audio. Next, pitch saliency and continuity constraints are applied to estimate the predominant melodic pitch. The best of pitch detection methods achieve no more than 80% accuracy on polyphonic audio. An important factor limiting the accuracy is the fixed choice of analysis parameters, which ideally should be matched to the characteristics of the audio such as the pitch range of the singer and the rate of variation of pitch. In the regions of rapid pitch modulation, characteristic of Indian classical singing, shorter analysis windows serve better to estimate the vocal harmonic frequencies and amplitudes. Hence for better pitch detection accuracy, it is necessary to adapt the window length to the signal characteristics. This is achieved automatically by the maximization of a signal sparsity measure computed at each analysis instance (every 10 ms) for local pitch detection (Rao et al., 2012). Finally, it is necessary to identify the vocal regions in the overall tracked pitch. This is achieved by using the peculiar characteristics of Hindustani music where the vocal segments are easily discriminated from the instrumental pitches due to the different temporal dynamics (Rao et al., 2011).

3.2.2 Motif Candidate Selection

Motivated by the characteristic of the *mukhda*, namely its alignment with the *sam* stroke of the rhythm cycle which the artiste pays great importance to achieve, the search algorithm starts by restricting candidate motifs to those that match rhythmically. This can be achieved via the automatic detection of the beat instants in the audio. In the spectrogram of Figure 3.1, the tabla strokes corresponding to the beats of the *tala* cycle are visible as vertical impulsive onsets. While the *sam* stroke itself is not particularly distinctive, the *dha* strokes (including the *sam*) can be detected as the highest onsets in the combined energies of two frequency bands: [5000, 8000] and [0, 1000] Hz. The former band is relatively free of interference from vocal partials, while the latter band captures the low frequency partial of the *dha* stroke. The filtered output power is subjected to a first-order difference and then half-wave rectified. Spurious peaks are removed by a local threshold. The consistency of the spacing of detected onsets with the known average tempo is considered further to identify the largest peaks as the *dha* stroke onsets. All audio segments whose alignment around a detected onset matches that of the *mukhda* are treated as potential candidates for *mukhda* detection. The extracted segment extends from the instant (*sam*-t1) to (*sam*+t2) where t1 and t2 are nominal values (number of beats in the 16-beat cycle) chosen based on the reference *mukhda* instance. The beat duration at a particular time instance is computed from the duration between the sub-cycle beats in the vicinity. For *tintal*, since there are 4 beats in a sub-cycle, the duration of a beat will be $\frac{sub\text{-}cycle\ duration}{4}$. Such a data representation is inherently robust to the slow tempo variations that occur during the concert.

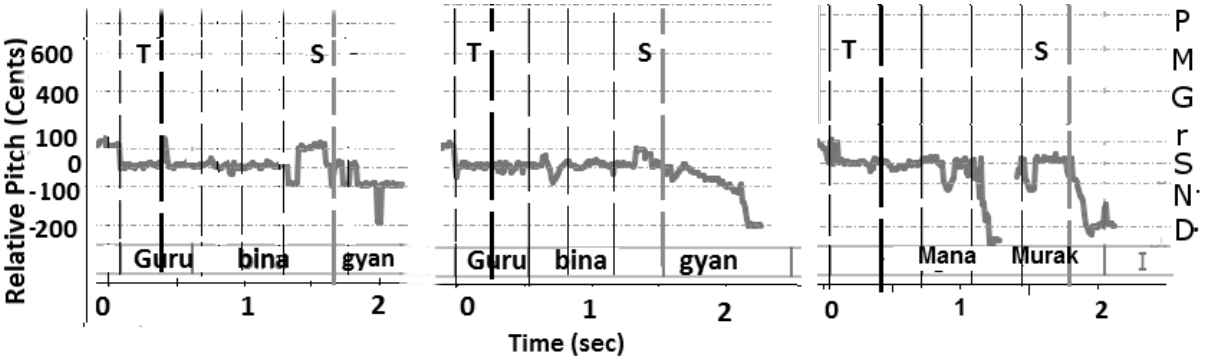


Figure 3.3: Two positive and one negative phrases of *Guru Bina Gyan* S:*sam* T:non-*sam* sub-cycle beat

The sequence of pitch values (in cents) obtained across the extracted candidate audio segment is a time series representation that is used further for similarity matching with a reference

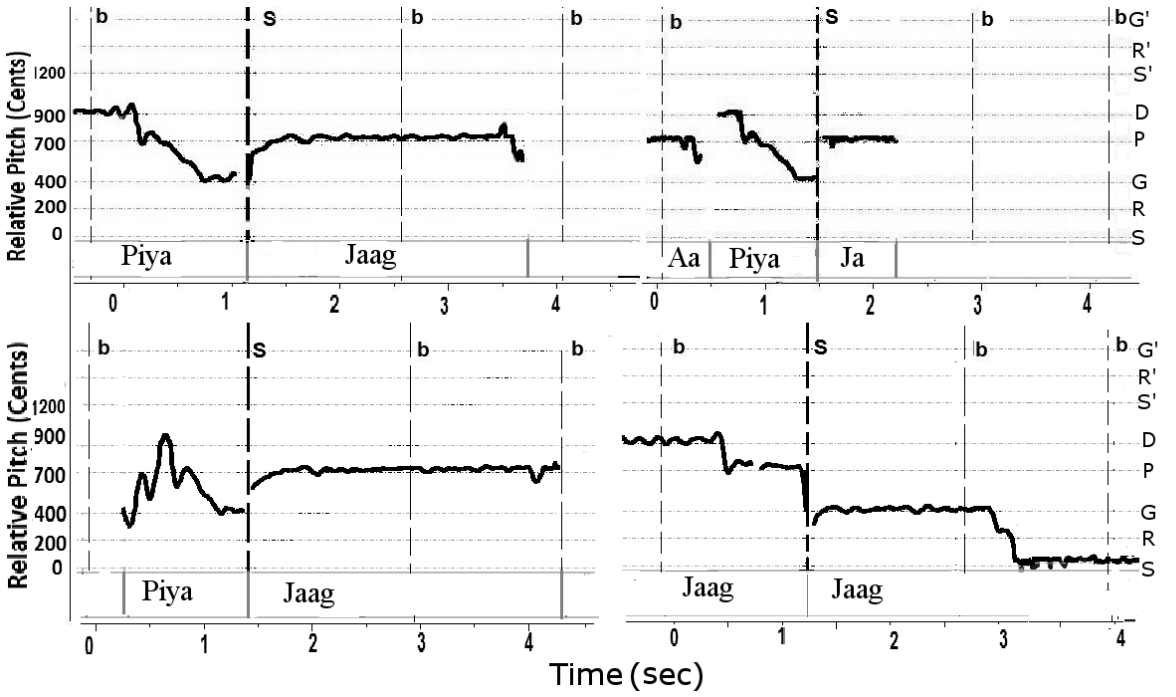


Figure 3.4: Three positive and one negative (bottom right) phrases of Piya Jaag S:*sam* b: beat

time series that is similarly obtained. Cents is used to express music intervals on a logarithmic scale. Figures 3.3 and 3.4 depict the pitch contours of a few candidate segments showing examples of the melodic and timing variability across *mukhda* realization within concerts. We observe that there are prominent differences in the melodic pitch contour, both in terms of fine pitch variation as well as timing. The note (*swara*) sequence of the *Guru Bina Gyan* phrase is seen to be [*S, S, N, R, N, D*]. However, since the word *Gyan* is often left unsung by the artiste, the *sam* itself serves as the right limit (i.e. t1=5, t2=0) of the *mukhda* in our task. The *swaras* corresponding to *Piya Jaag* are [*D, P, G, P*]. Here t1=1 and t2=2 were applied to delimit the *mukhda*. Any pitch gaps within the boundaries correspond to pauses. These are filled by linear interpolation or extrapolation of neighboring pitch values before similarity matching.

3.2.3 Similarity Modeling

Due to the beat-based method of candidate extraction, the segments tend to be of different absolute durations depending on local tempo variations. Also, singing expressiveness manifests itself in timing changes that can affect the total duration of the sung phrase. The sequence of pitch values obtained at 10 ms intervals throughout the candidate audio segment can be viewed as a non-uniform length time-series representation. We explore two distinct similarity measures

for non-uniform length temporal sequences; Symbolic Aggregate approXimation (SAX) and Dynamic Time Warping (DTW).

SAX

SAX reduces a time series of length n to a string of length w ($w < n$) through piecewise aggregate approximation. Piecewise aggregate approximation has been used to obtain dimension-reduced uniform length time-series for motif discovery in bio-informatics (Lin et al., 2003). We apply this method to convert a non-uniform length time series of pitch in cents, computed every 10 ms, to a uniform length, dimension-reduced sequence of symbols. The string length w is varied to determine the optimum dimension of the data representation. A given time-series ($C = c_1, c_2, ...c_n$) is aggregated into w uniform length segments ($\bar{C} = \bar{c}_1, \bar{c}_2, ...\bar{c}_w$) each represented by the averaged value of the segment. $\bar{c}_i$ is computed as

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \tag{3.1}$$

The real-valued pitch in cents is retained as such, but we also consider quantizing pitch to the nearest semitone. Since the tonic frequency is singer dependent in Indian music, the semitone grid is anchored on the most prominent peak of an overall pitch histogram derived from the vocal pitch track across the test audio. Since our present work is confined to within-concert matching, a tonic detection error is inconsequential. Next, the Euclidean distance between the two W-length sequences, the reference and the candidate, is used as a similarity measure.

DTW

Another widely used method to compare real-valued time series related to each other through, possibly, nonlinear time-scaling, is the dynamic time-warping distance measure (Berndt and Clifford, 1994). DTW effectively handles local stretches and shrinks to compute meaningful distance between time series sequences (Fu et al., 2008). Given two time series sequences ($x_1, x_2, ...x_n$ and $y_1, y_2, ...y_m$), DTW finds the optimal warping path between them and the accumulated cost matrix can be computed as

$$DTW(i,j) = cost(x_i, y_j) + min \left\{ \begin{array}{c} DTW(i-1,j) \\ DTW(i-1,j-1) \\ DTW(i,j-1) \end{array} \right\} \tag{3.2}$$

where $cost(x_i, y_j)$ is the local distance measure between x_j and y_j (Muller, 2007). The distance between the so aligned reference and candidate phrases is used as the similarity measure. Pathological warpings are avoided by incorporating the Sakoe-Chiba constraint (Sakoe and Chiba, 1978) on the width of a diagonal band in the DTW path matrix. Sakoe-Chiba band defines a global constraint restricting the warping to lie within a band surrounding the diagonal. The absolute difference in cents between pitch values is used as the local distance in the DTW path optimization. Any absolute difference within 25 cents (i.e. a quarter tone) is rounded down to 0 cents. This is found to help reduce the influence of practically imperceptible pitch differences on the warping path and therefore any unnecessary stretching of the path.

3.3 Experiments and Results

Given the database described in Table 1, we evaluate the different data representations and similarity measures on within-concert and across singer-concert motif detection tasks. Each candidate phrase extracted from the detected onsets as presented in Section 3.2 is labeled positive or negative depending on whether or not it is the actual motif (i.e. *mukhda* phrase). Table 1 shows the number of such phrase segments available for the evaluation of the motif detection methods. To maximize the use of the available annotated data, each labeled motif is considered as the reference once with all other motifs serving as positive tokens and the remaining candidates as negative tokens. Thus, the *Piya Jaag* motif detection task can be evaluated on 44×43 = 1892 positive pairs and 44×176 = 7744 negative pairs (i.e. each positive with all negatives). Table 2 summarizes the experiments. The Experiment A considers motif detection from within the *Guru Bina* recording of Bhimsen Joshi given a reference template from the same recording. Similarly, the Experiments B, C and D consider the within-concert detection as specified in Table 2. The Experiment E uses the positive tokens of *Guru Bina* by BJ to detect the *mukhda* in the same *bandish* concert by a different vocalist, AC. As it turns out, the two male singers are tuned to the same tonic. In each experiment, the rate of false alarms for a given hit rate (correct detections) is computed for each combination of similarity model and data representation.

The similarity measures include SAX and DTW. The data representations chosen for the study are either the continuous pitch values (i.e. 1200 cents per octave) indicated by q1200, or the quantized versions (12 semitones per octave on an equi-tempered scale) indicated by q12.

Expt	Bandish	Singer	#Phrases	
			POS	NEG
A	Guru Bina	BJ	156	715
B	Guru Bina	AC	1056	9735
C	Jana na na na	BJ	272	1649
D	Piya Jaag	KA	1892	7744
E	Guru Bina	BJ Vs AC	429	3835

Table 3.2: Description of experiments with number of positive and negative phrase candidates available in each performance

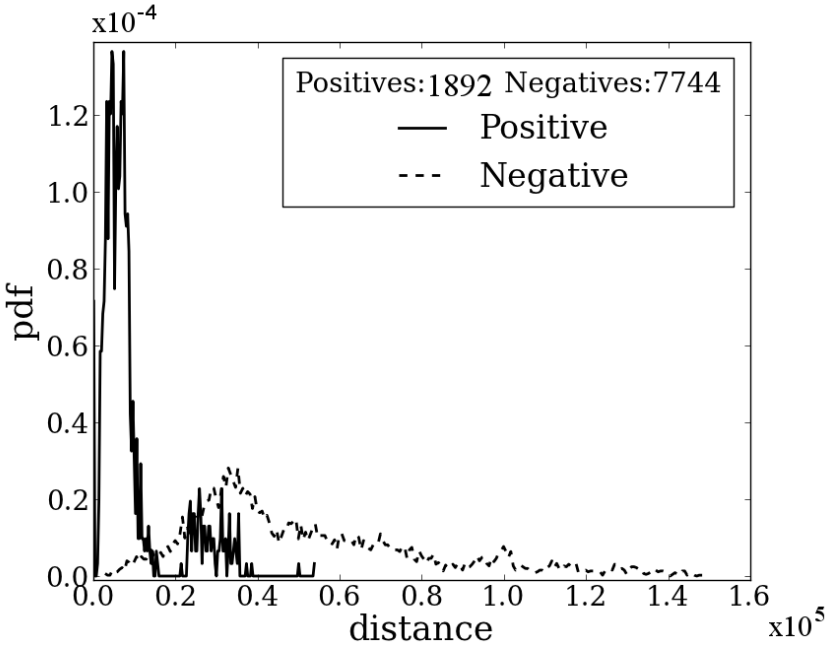


Figure 3.5: DTW distances distribution for Piya Jaag recording

Figure 3.5 shows an example of the distribution of distances for positive-positive pairs and positive-negative pairs. The recording is *Piya Jaag* (Experiment D) evaluated with DTW-q1200. We observe that the distances between the positive phrases cluster closely relative to the distances between the positive-negative phrase pairs. There is a limited overlap between

the two distributions. That the spread of the negative distances is relatively wide indicates the robustness of the distance measure in terms of its discrimination of melodic shapes. We also note the presence of a small isolated cluster of positive distances. A closer examination revealed that this stemmed from the wide timing variability across *Piya Jaag* phrases with its particularly slow tempo. Thus there were at least two distinct behaviours within the set of positive phrases. The inter-phrase distances between the longer duration phrases tended to be lower than the distances involving shorter duration phrases. Figure 3.6 shows the ROC (hit rate versus false alarm rate) derived from the distributions of Figure 3.5 by varying the decision threshold. We observe two bends in the curves, consistent with the bimodal shape of the pdf.

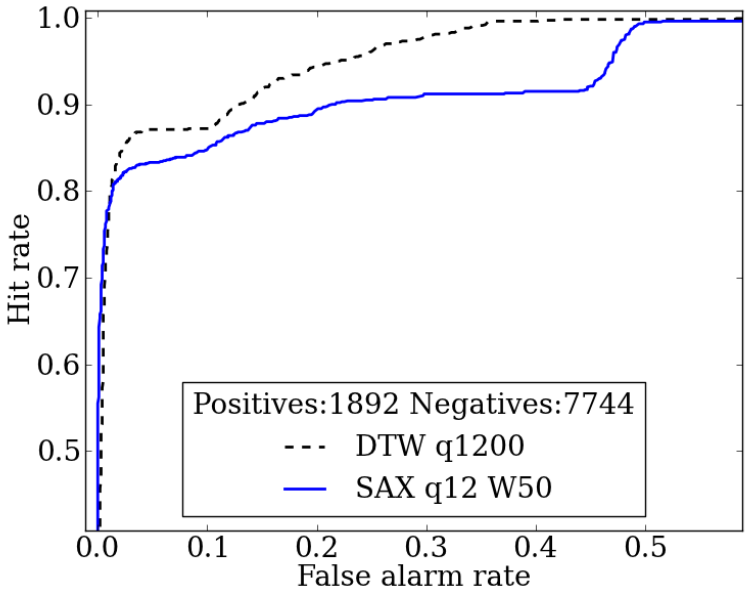


Figure 3.6: ROC curves for Piya Jaag distribution

Table 3.3 summarizes the classification results across the experiments in terms of false alarm rate (FA) for a fixed HR chosen near the knee of the ROCs of the corresponding data. Given that the extracted candidate phrases have durations varying in the range of 2-4 sec (200-400 length string), we vary the SAX string length around W=50 (corresponding to the aggregation of 4-8 samples). Preliminary experiments revealed that, W substantially lower than this value(i.e. more averaging) led to worsened performance. We note that the performance of SAX improves with pitch quantization at a fixed string length of 50. Increasing or decreasing the string length around this does not improve performance on the whole. The DTW system performs substantially better than SAX in terms of reducing the FA at fixed hit rate. As in the case

Method	Experiment A		Experiment B		Experiment C		Experiment D		Experiment E	
	HR	FA	HR	FA	HR	FA	HR	FA	HR	FA
SAX-q1200-W50	1	.096	.94	.019	.86	.239	.87	.130	.94	.035
SAX-q12-W40	1	.084	.94	.016	.86	.231	.87	.135	.94	.024
SAX-q12-W50	1	.071	.94	.015	.86	.216	.87	.124	.94	.029
SAX-q12-W60	1	.091	.94	.014	.86	.210	.87	.133	.94	.023
DTW-q1200	1	.044	.94	.007	.86	.044	.87	.032	.94	.015
DTW-q12	1	.053	.94	.008	.86	.042	.87	.025	.94	.014

Table 3.3: Performance of SAX and DTW motif detection under different configurations. WX = SAX string dimension is X; qY= quantized pitch levels per octave; HR = hit rate; FA = number of false alarms

of SAX, pitch quantization helps to improve performance further in some cases. That DTW does relatively better indicates that non-uniform time warping is essential to achieve the needed alignment between phrases before distance computation. This is consistent with what is known about the tradition where the essential melodic sequence of the *mukhda* phrase is strongly adhered to by the singer while metrical alignment is focused only on getting to the specific syllable onset (e.g. *Gyan* in Figure 3.3 and *Jaag* in Figure 3.4) on the first beat of the cycle (the *sam*).

Further, comparing the Experiments E and B as a case of between-concert to within-concert performance of the motif detection methods, we see that the FA is somewhat higher in Experiment E which involves a reference motif from the concert of the same *bandish* by a different artiste. This is consistent with the anticipated higher variability in motif contour across artistes.

3.4 Summary

The methods are investigated in the context of detecting the signature phrase of Hindustani vocal music compositions within and across performances. *Mukhda* identification being one of the fundamental problems in Hindustani MIR realm, makes it relevant. Further, it has pedagogical applications. The processing of the polyphonic audio signals needed to achieve a suitable data representation was presented. Musical knowledge related to the metrical relation between the *mukhda* motif and the underlying rhythmic structure was exploited to achieve a reduced search space, using available similarity measures, and possibly more robust detection. Similarity mea-

sures traditionally used in time-series matching have been shown to perform well in the context of melodic motif detection in the improvised *bandish* of Hindustani vocal concert recordings. We observed that, DTW gives a relative better accuracy over SAX. Non-uniform time-warping of DTW helps. Tempo normalization alone is insufficient to capture the melodic invariance of the motif. Also we observed that, quantization to 12 semitones does not reduce the accuracy. The *mukhda* context considered in this work is relevant in both Hindustani and Carnatic vocal music (in the *bol-alap* and *niraval*, respectively). In the future, extension of this dataset with performances having highly varying *mukhda* instances, will enable to research on more robust methods.

Chapter 4

Pakad Detection

The essence of a raga is captured in a set of melodic phrases known as the *pakad* (catch phrases) widely used in both compositions in the raga, and in improvisation where the identity of the raga is revealed by the artiste through the use of the ragas melodic motifs. The set of phrases form the building blocks for melodic improvisation by collectively embodying the ragas melodic grammar and thus defining a ragas personality (Raja, 2005). The identity of a *pakad* lies in the *swara* combination, the intonation of the *swaras* and the ornamentations. Preserving the characteristics of a *pakad*, the artiste is granted the flexibility to improvise it. The task of *pakad* identification is to discover instances of a *pakad* from a performance, despite the variations present within the *pakad* instances.

In view of the central role played by raga-characteristic phrases in the performance of both the Indian classical traditions, computational methods to detect specific phrases in audio recordings can have important applications. Raga-based retrieval of music from audio archives can benefit from automatic phrase detection, where the phrases are selected from a dictionary of characteristic phrases corresponding to each raga (Chakravorty et al., 1989). Given that the identity of a raga-characteristic phrase is captured by the *swaras* and *gamakas* that constitute it, a melodic representation of the phrase would be crucial in any automatic classification task. Phrase recognition can help in the automatic music transcription of Indian classical music which is notoriously difficult due to its interpretive nature. Such phrase-level labeling of audio can be valuable in musicological research apart from providing for an enriched listening experience for music students.

In this work, we restrict ourselves to the problem of classifying pre-segmented phrases in concert audio recordings within a closed set of characteristic phrases of one or more ragas. The

primary focus is on the similarity computation method to enable better distinction between the phrases from different *pakad* classes. We investigate on, how learning the global constraints for DTW can be beneficial to this task. As discussed in Section 3 of the previous chapter, continuous pitch contours corresponding to the raga-characteristic phrases are likely to serve cognitively as the best units of melody. Here also we utilize the predominant pitch detection algorithm to extract the pitch contour (Rao and Rao, 2010).

This chapter is organized as follows. To give a better idea about the task, we start with description on dataset and annotation in Section 4.1. Section 4.2.1 discusses the pre-processing of the audio content to extract the melodic representation of the phrases. Section 4.3 presents a detailed analysis of the phrase-level characteristics of the *pakads* considered for the study. The discussion on similarity computation in Section 4.4 describes the core idea including clustering of the training phrases and learning the DTW global constraints. Section 4.5 discusses the experiments and results.

4.1 Dataset and Annotation

Concert audio recordings were used to put together the raga motifs database. The dataset is assembled from the *bada khayal* sections where the *vistar* is the improvised segment that occurs within the constraints of the rhythmic cycle framework of the *bandish*. Sequences of raga-characteristic phrases appear between occurrences of the *mukhda* (or refrain) of the composition. Several general properties of the melodic motifs are illustrated through examples from these datasets.

For the *pakad* detection task, we choose the *pakads* of the raga Alhaiya Bilawal, a commonly performed raga of the Bilawal group, which includes ragas based on the major scale (Rao et al., 1999). It is considered to be complex in its phraseology and is associated with a sombre mood. While its notes include all the notes of the Western major scale, it has additionally the *komal Ni* (*n*) in the descent (*avaroha*). Further *Ma* is omitted from the ascent. The typical phrases used for raga elaboration in a performance appear in Table 4.1. A specific phrase may appear in the *bandish* itself, or in the *bol-alap* and *bol-taan* (improvised segments). It may be uttered using the words or syllables of the *bandish* or in *aakar* (melismatic singing on the syllable /a/). What is invariant about the *calana* is its melodic form which may be described as a particular-shaped pitch trajectory through the nominal notes (swaras) in Table 4.1. Raga Kafi,

Raga Characteristics	**Alhaiya Bilawal**	**Kafi**
Tone Material	S R G m P D n N	S R g m P D n
Characteristic Phrases	G∼ R G /P (GRGP) D∼ n D \P (DnDP) D \G G m R G P m G	R g R m m P g- m P m P D m n \P g R S n \P g R
Comments	'n' is used only in the descent, and always in between the two 'D'-s as D n D P	Movements are flexible and allow for melodic elaboration

Table 4.1: Raga descriptions adapted from (Rao et al., 1999). The characteristic phrases are provided in the reference in enhanced notation including ornamentation. The prescriptive notation for the phrases used for the present study appears in parentheses

whose description also appears in Table 4.1, is used in this work primarily as an anti-corpus, i.e. to provide examples of note sequences that match the prescriptive notation of a chosen *pakad* of Alhaiya Bilawal but in a different raga context (and hence are not expected to match the melodic shape, or intonation). *DnDP* is a *pakad* in Kafi, but has different characteristics compared to *DnDP* in Alhaiya Bilawal.

A total of eight selected audio recordings of Raga Alhaiya Bilawal and one recording of Raga Kafi, by eminent Hindustani vocalists, from commercial CD and NCPA AUTRIM archive for Music in Motion (Rao, 2013) have been used for the study. The common accompanying instruments were tanpura (drone), tabla and harmonium, except one having sarangi in place of harmonium. The concert sections of *bandish* with its *vistar* (only non-*taan* section) spanning slow (*vilambit*) to medium (*madhya*) tempos. Although the size of the database is limited, it has been designed to present a challenging scenario by including related phrases in raga Alhaiya Bilawal that share a common *nyas* (focal or ending). From Table 4.1, these are *DnDP* and *GRGP*. Additionally, the phrase *mnDP* which occurs in the *mukhda* in several of the recordings is also used due to its similarity to *DnDP* in terms of shared *swaras*. Finally, the melodic segments corresponding to the *DnDP* sequence from raga Kafi are included due to their shared prescriptive notation with the raga-characteristic phrase in Alhaiya Bilawal. The phrases were labeled by a musician (and later validated by another) using the PRAAT interface (Boersma

and Weenink, 1992). A steady note or rest generally cues a phrase listened for the occurrence of the P-*nyas* ending phrases in Table 4.1. Every recognized instance was marked with the corresponding phrase name in corresponding location of the PRAAT textgrid layer. Alhaiya Bilawal *DnDP* phrase was clearly distinguished from the *DnDP* sequence segments by the phrase intonation and sometimes, preceding context. These sequence instances share the same notation, but does not have the characteristics of the *pakad DnDP*. Note that there was no effort to mark precise phrase boundaries; rather the phrase was just coarsely delimited via markers on the waveform. The actual phrase boundaries were refined via automatic segmentation described later. A count of the phrases of each category appears in Table 4.2.

SongID	**Artiste**	**Tala**	**Laya**	**Bandish**	**Tempo (bpm)**	**dur. (min)**	**#Phrases**			
							DnDP		*mnDP*	*GRGP*
							Char.	Seq		
AB	Ashwini Bhide	Tintal	Madhya	Kavana Batariyaa	128	8.85	13	2	31	5
MA	Manjiri Asanare	Tintal	Vilambit	Dainyaa Kaahaan	33	6.9	12	1	13	6
SS	Shruti Sadolikar	Tintal	Madhya	Kavana Batariyaa	150	4.15	3	0	14	3
ARK	Abdul Rashid Khan	Jhaptal	Madhya	Kahe Ko Garabh	87	11.9	44	0	0	14
DV	Dattatreya Velankar	Tintal	Vilambit	Dainyaa Kaahaan	35	18.3	14	4	4	9
JA	Jasraj	Ektal	Vilambit	Dainyaa Kaahaan	13	22.25	19	18	0	29
AK1	Aslam Khan	Jhumra	Vilambit	Mangta Hoon Tere	19	8.06	10	0	8	6
AK2	Aslam Khan	Jhaptal	Madhya	E Ha Jashoda	112	5.7	7	0	0	3
AC	Ajoy Chakrabarty	Jhumra	Vilambit	Jago Man Laago	24	30.3	—	27	0	—
			Total no. of Phrases				122	52	70	75

Table 4.2: Description of database with phrase counts in the musicians transcription; all concerts are in raga Alhaiya Bilawal except the last (AC) in raga Kafi. Char. = characteristic of the raga; Seq. = note sequence

4.2 Audio Processing

In this section, we discuss the processing of the audio signal to extract the melodic representation of the phrase. The continuous pitch versus time is a complete representation of the melodic shape of the phrase, assuming that volume and timbre dynamics do not play a role in motif recognition by listeners.

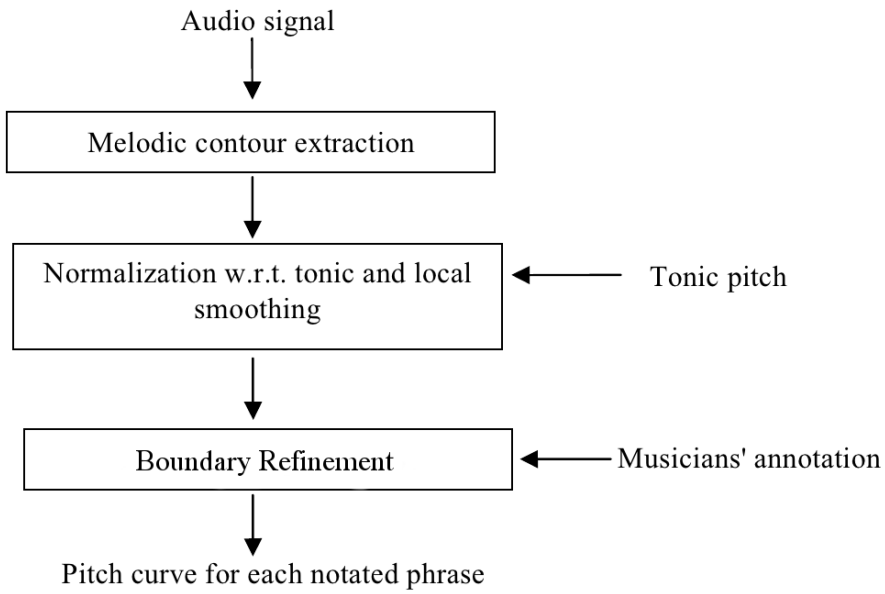


Figure 4.1: Block diagram for audio processing

4.2.1 Audio Processing Stages

Melodic Contour Extraction

The singing voice usually dominates over other instruments in a vocal concert performance in terms of its volume and continuity over relatively large temporal extents although the accompaniment of tabla and other pitched instruments such as the drone and harmonium or violin is always present. Melody extraction is carried out by predominant-F0 detection methods (Rao and Rao, 2010) as explained in the previous chapter, Section 3.2.1. These methods exploit the local salience of the melodic pitch as well as smoothness and continuity over time to provide a pitch estimate and voicing decision every frame. We use frame durations of 10 ms giving us a continuous pitch curve (Hz versus time), sampled at 10 ms intervals, corresponding to the melody.

Normalization and Local Smoothing

The *swara* identity refers to the pitch interval with respect to the artiste-selected tonic. In order to compare the phrase pitch curves across artistes and concerts, it is necessary to normalize the pitches with respect to the chosen tonic of the concert. Thus the pitches are represented in cents with respect to the detected tonic of the performance (Salamon et al., 2012). The pitch curve next is subjected to simple 3-point local averaging to eliminate spurious perturbations that may arise from pitch detection errors.

Segment boundary refinement

Since the scope of the present work is restricted to classification of segmented phrases, we use the musicians' labeling to extract the pitch curve segments corresponding to the phrases of interest. Since the musicians' labeling is carried out relatively coarsely on the waveform in the course of listening, it is necessary to refine the segment boundaries in order to create pitch segments for the training and testing sets of similarity computation especially in the case of exemplar based matching. Thus phrase segmentation is carried out on the Hindustani audio melodic contours in a semi-automatic manner by detecting the onset and offset of the starting and ending notes respectively. An onset or offset of a *swara* is reliably detected by hysteresis thresholding with thresholds of 50 and 20 cents within the nominal pitch value. Figure 4.2 shows the *DnDP* phrase segment where the phrase boundaries are eventually marked at offset of the *n* (descending from *S*), and onset of the P-*nyas*.

The output of the audio processing block is the set of segmented pitch curves that correspond to the phrases of interest as shown in Table 4.2. Thus each phrase is represented by a tonic-normalized cents versus time continuous pitch curve. Figure 4.3 shows examples of the pitch curves. These are of varying duration depending on the duration of the phrase in the audio.

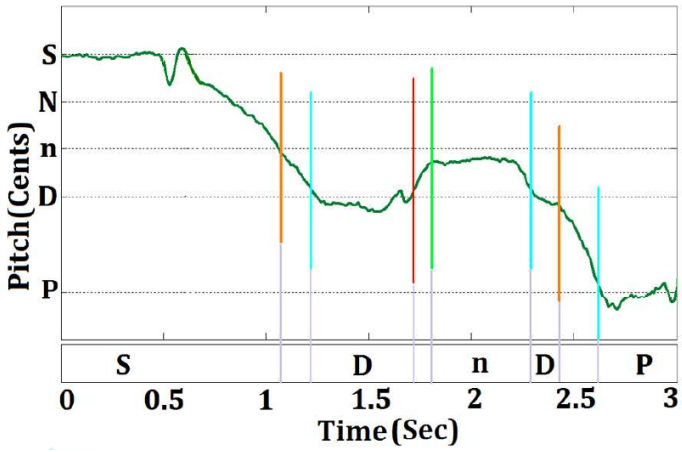


Figure 4.2: Illustrating *swara* onsets and offsets (colored vertical bars) in a *DnDP* phrase pitch curve. Orange/red: exiting a *swara* by descending/ascending; Green/blue: approaching a *swara* from below/above. Phrase boundaries are selected from these instants based on the starting and ending *swara* s of the phrase

4.3 Phrase-level pitch curve characteristics

In this section, we present a detailed analysis of the characteristics of the *pakads*; *DnDP* and *mnDP* in raga Alhaiya Bilawal. Though this analysis is limited to only two different *pakads*, this analysis aims at getting insights into the variabilities and invariabilities of a *pakad*. We also analyze intra-phrase class similarities through distance distribution of the *pakad* instances from different performances of the same raga. Here the identity of a *pakad* in terms of its invariabilities is also analyzed, comparing the distance distribution of *DnDP pakad* instances with distance distribution of sequence *DnDP* instances which are not considered as *pakad*.

Figure 4.3 (Image 1) shows some representative pitch contours for *DnDP* phrases in various melodic contexts selected from different concerts in our dataset (Table 4.2). The contexts typically correspond to the two possibilities: approach from higher and approach from lower *swara*. The vertical lines mark the rhythmic beat (*matra*) locations whenever these were found in the time region covered in the figure. We consider the phrase duration, indicated by the dark vertical bars, as spanning from D-onset or m-onset (or rather, from the offset of the preceding n through which *swara* the phrase is approached) to P-onset. The final *P* is a resting note and therefore of unpredictable and highly varying duration. From the spacing between beat instant markers in Figure 4.3, we note that the MA concert tempo is low relative to the others. However the phrase durations do not appear to scale in the same proportion. It was noted that across the

concerts, tempi span a large range (as seen in Table 4.2) while the maximum duration of the *DnDP* phrase in any concert ranges only between 1.1 to 2.8 sec with considerable variation within the concert. Further, any duration variations of sub-segments are not linearly related. For example, it is observed that the n-duration is practically fixed while duration changes are absorbed by the D *swara* on either side. There was no observable dependence of phrase intonation on the *tala*. Apart from these and other observations from Figure 4.3, we note that the raga Kafi phrases (in which raga, *DnDP* is not a characteristic phrase but merely an incidental sequence of notes) display a greater variability in phrase intonation while conforming to the prescriptive notation of *DnDP* (Rao et al., 2013). We investigated this in detail through 2 different experiments with the data mentioned in Table 4.2.

We observe the similarity in melodic shape across realizations of a given phrase in the Alhaiya Bilawal raga. Prominent differences are obvious too, such as the presence or absence of *n* as a touch note (*kan*) in the final *DP* transition in *DnDP* and varying extents of oscillation on the first D. The similar comments apply to the different instances of *mnDP* shown in Figure 4.3. Variations within the phrase class may be attributed to the flexibility accorded by the raga grammar in improvisation. Consistent with musicological theory on *khayal* music at slow and medium tempi, (i) there is no observable dependence of phrase duration upon beat duration, (ii) relative note durations are not necessarily maintained across tempi, and (iii) the note onsets do not necessarily align with beat instants except for the *nyas*, considered an important note in the raga.

4.3.1 Intra-phrase-class similarity

The prominent similarity characteristics are preserved among instances of a *pakad* without being dependent on the performance. We validate this invariance across concerts and artistes with the melodic pitch contours extracted from audio recordings. The ability to distinguish a selected raga-characteristic phrase from other phrases of the same raga and from the similarly notated phrase in a different raga is examined.

For this experiment we compare the pitch curve similarity across *DnDP* phrases of the raga-characteristic class with that of that across the non-characteristic phrase class. The DTW distance measure is applied to all pairs created from two distinct *DnDP* phrases drawn from the same class. If implemented for each concert audio, this could tell us something about the variability of the *DnDP* phrase in that concert. To compensate for the shorter durations of some

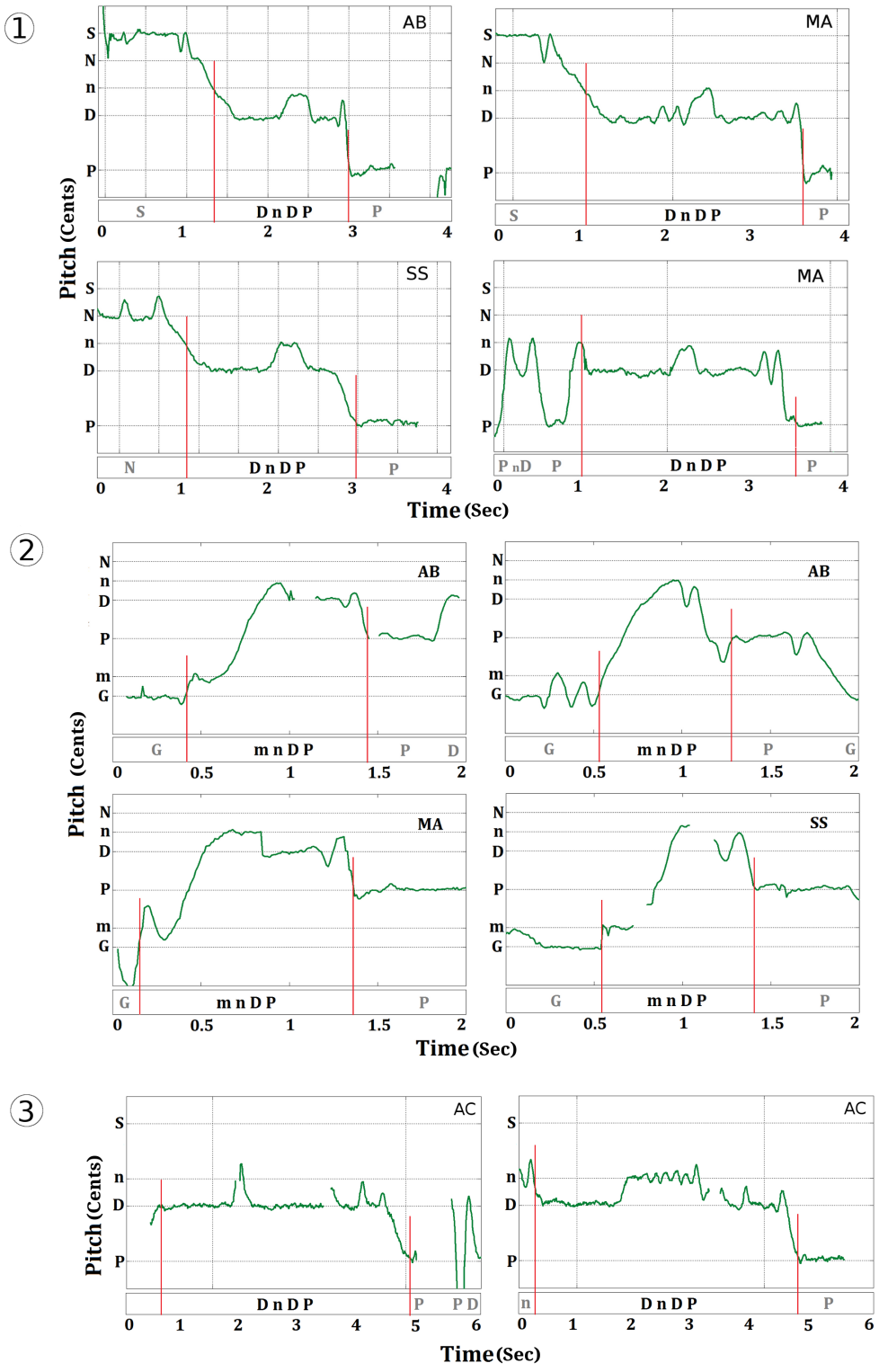
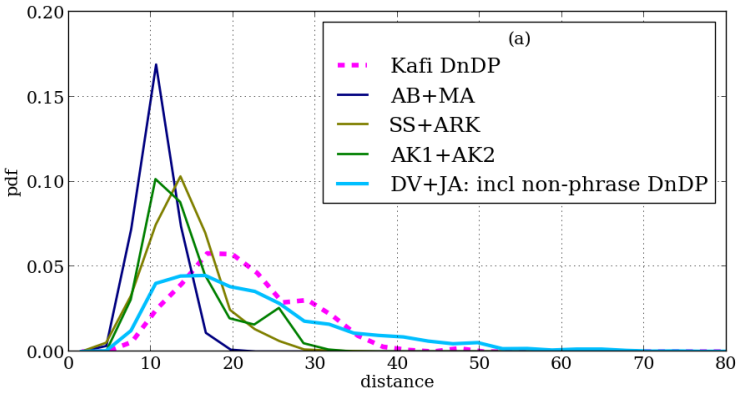
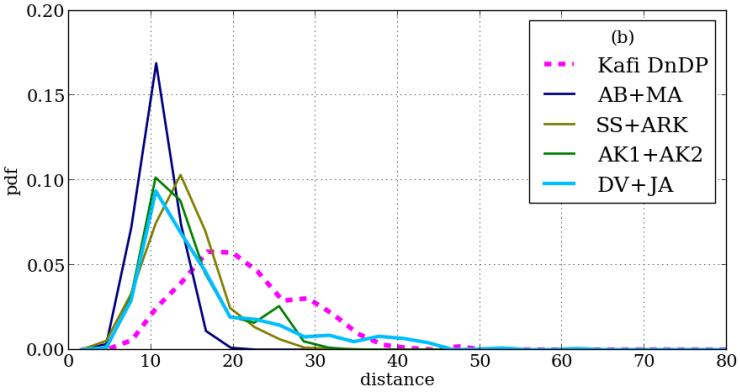


Figure 4.3: Pitch contours (cents vs time) of different phrases in various melodic contexts by different artistes. Horizontal lines mark *swara* positions. Thin vertical lines mark beat instants. Thick lines mark the phrase boundaries for similarity matching; 1.Alhaiya Bilawal *DnDP*, 2.Alhaiya Bilawal *mnDP*, 3.Kafi *DnDP*

of the concerts however, we group the 8 Alhaiya Bilawal concerts in Table 4.2 into 4 sets. The resulting distribution of pair-wise phrase distances for each set is shown in Figure 4.4(a). The number of pairs is given by N(N-1) where N is the count of *DnDP* phrases in that set (e.g. N=17 for AK1+AK2). Also shown is the distance distribution created out of *DnDP* phrase pairs from raga Kafi.



(a)



(b)

Figure 4.4: Intra-phrase-class distance distributions for the different concerts listed in Table 4.2 (a) All *DnDP* sequences included (b) Non-characteristic *DnDP* excluded in the Alhaiya Bilawal concerts.

Figure 4.4(a) shows that the intra-phrase-class distances in all Alhaiya Bilawal concerts (except DV+JA) are narrowly dispersed about a mean value close to 12.0, indicating the low variability in the intonation of the raga-characteristic phrase across the concert. In contrast, the Kafi raga distribution has a mean near 20.0 and higher standard deviation, implying greater variability in *DnDP* intonation. These observations are consistent with musicological knowl-

edge about the strictness that applies to the melodic shape of raga-characteristic phrases as opposed to that of non-characteristic phrases. The DV+JA concerts show a relatively greater spread due to the presence of non-characteristic *DnDP* (as indicated in Table 4.2). When these non-characteristic phrases are eliminated from the computed distances, we have the more concentrated distribution for DV+JA in Figure 4.4(b). Thus we see that the mean and spread of the inter-phrase distance distribution clearly capture the raga characteristics with respect to the given phrase. This experiment throws light to the fact that, inspite of the flexibility available to artiste in rendering a raga-characteristic phrase, distinctiveness w.r.t other ragas is retained.

4.4 Similarity Computation

We present an improved DTW based pattern matching method to achieve the classification of the test melodic segments, obtained as described in the Section 4.2.1. Here we learn the global constraints for DTW instead of a fixed-band constraint like Sakoe-Chiba. The learned global constraints capture the variabilities and invariabilities of a *pakad* class through variable constraining. For each *pakad* class, the constraints are learned with the phrases from a subset of performances in the dataset taken as training dataset. Also for exemplar-based matching, reference templates for each phrase class of interest are automatically identified from the training set of phrases. A DTW based distance measure is computed between the test segment and each of the reference templates. The detected phrase class is that of the reference template that achieves the lowest distance, provided the distance is below a pre-decided threshold. DTW distance computation combines a local cost with a transition cost, possibly under certain constraints, both of which must be defined meaningfully in the context of our task of melodic matching.

4.4.1 Classification of test segment

The test pitch curve obtained from the audio processing previously described is prepared for DTW distance computation with respect to the similarly processed reference template as shown in the block diagram of Figure 4.5. The phrases are of varying duration, and a normalization of the distance measure is achieved by interpolating the pitch curves to the fixed duration of the reference template that it is being compared with. The fixed duration of a reference template is the average duration of the phrases that it represents. Before this, the short gaps within the pitch curve arising from unvoiced sounds or singing pauses are linearly interpolated using pitch

values that neighbor the silence region up to 30 ms. Zeros are padded at both the ends of the phrases to absorb any boundary frame mismatches which has adverse effect on DTW matching.

To account for the occurrence of octave-transposed versions of phrases with respect to the reference phrase, transposition of +1 and -1 octave creating 3 versions of the test candidate for the similarity matching are computed. We also investigate the quantization of pitch in the melodic representation. 12-semitone quantization to an equitempered scale (with respect to the tonic) and 24 level quantization to quartertones are obtained for evaluation.

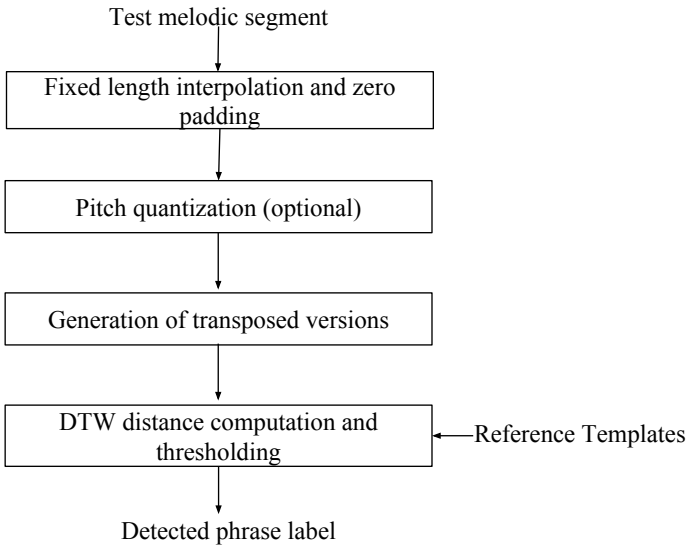


Figure 4.5: Steps in *pakad* classification

DTW distance between each of the template phrases and the transposed versions of the test pitch curve is computed. The template phrases comprise the codebook for the task and together represent the phrase classes of interest. For example, in our task we are interested in the recognition of *DnDP* and *mnDP* of raga Alhaiya Bilawal. The template phrases are therefore representative pitch curves drawn from each phrase class of interest in the labeled training dataset. Vector quantization (VQ), as presented in the next section on training, is used to obtain the codebook of phrases. The detected phrase class is that of the reference template that achieves the lowest distance, provided the distance is below a pre-decided threshold. From an informal examination of the labeled pitch curves, it was felt that two templates per phrase class would serve well to capture intra-class variabilities. Vector quantization also facilitates identifying the phrase instances belonging to different categories within the same *pakad* class. DTW global constraints are learned separately for each category that can potentially improve retrieval accuracy.

For a *pakad* class, while computing distance with a template phrase from a category, the constraint learned for that category is applied. Figure 4.6 shows the steps in constraint training and identification of template phrases for each category from the clustering of training phrases. Next section gives a detailed description about the clustering and constraint learning.

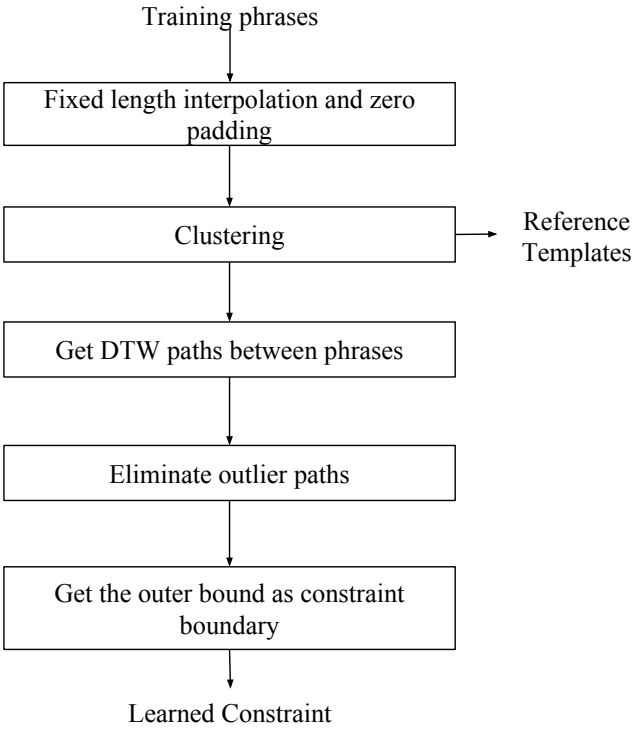


Figure 4.6: Steps for constraint training and identification of template phrases

4.4.2 Constraint Learning

Vector Quantization (Clustering)

As described, the clustering step segregates the phrases in the training set into categories. We use k-means algorithm to perform this clustering with k=2. k is fixed at 2 based on the above mentioned observation on intra-class categories within *DnDP* and *mnDP pakads*. A DTW distance measure is computed between fixed-length pitch curves. In each iteration of the k-means procedure, the centroid for each cluster is computed as the mean of corresponding pitches obtained after DTW time-aligning of each cluster member with the previous centroid. Thus we ensure that corresponding subsegments of the phrase are averaged.

Figures 4.7 and 4.8 show sample pitch curves from the clusters obtained after vector quantization of *DnDP* and *mnDP* phrases respectively. We observe that a prominent distinction be-

tween members of the *DnDP* phrase class is the presence or absence of the *n-kan* (touch note) just before the *P-nyas*. (Note that the *P-nyas swara* is not shown in the figure since it is not included in the segmented phrase as explained earlier). In the *mnDP* class, the clusters seem to be separated on the basis of the modulation extent of the initial *m swara*. Visually, the cluster centroids (not shown) are representative of phrase pitch curves of the corresponding cluster. Thus VQ of pitch curves serves well to capture the variations observed in phrase intonation. For each *pakad* class and for each category, the phrases closest to the cluster centroid are taken as template phrases. All the phrases in each cluster are considered for training the constraints for the corresponding category.

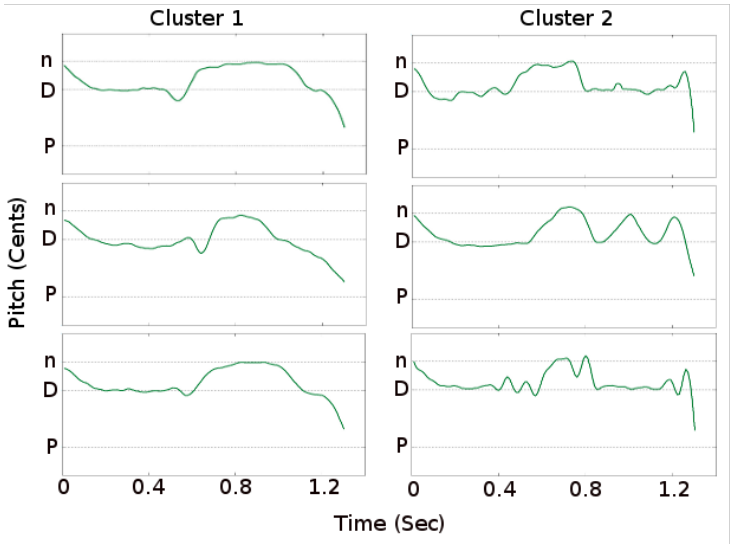


Figure 4.7: Examples from each of the two VQ clusters obtained for the *DnDP* instances from the AB and MA concerts (all phrases interpolated to uniform length of 1.3 seconds)

Constraint Learning

Global path constraints applied in DTW distance computation can restrict unusually low distances arising from pathological warping between unrelated phrases especially those that have one or more *swaras* in common. The global constraint should be wide enough to allow for the flexibility actually observed in the phrase intonation across artistes and concerts. As noted in Section 4.3, the elongation or compression observed in one instance of a phrase with respect to another is not uniform across the phrase. Certain sub-segments actually remain relatively constant in the course of phrase-level duration change. Thus it is expected that the ideal global constraint would be phrase dependent and varying in width across the phrase length. Apart

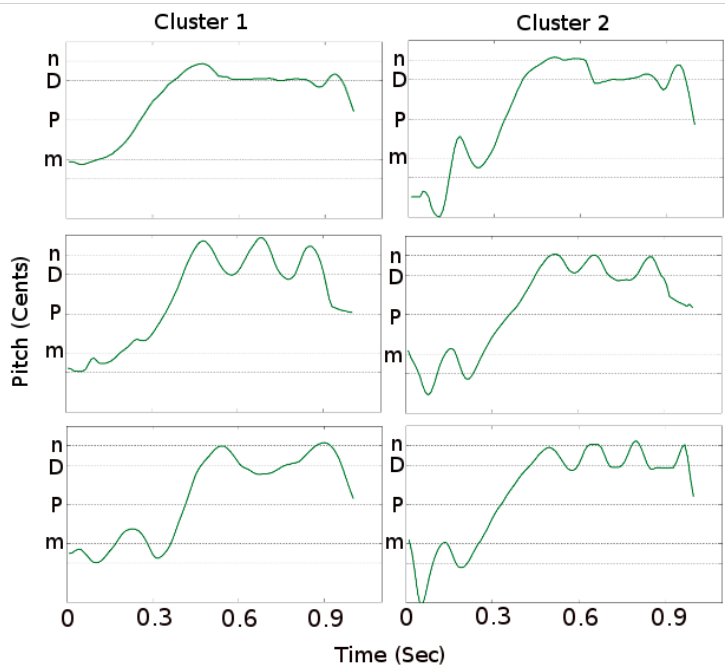


Figure 4.8: Examples from each of the two VQ clusters obtained for the *mnDP* instances from the AB and MA concerts (all phrases interpolated to uniform length of 1 second)

from the global path constraint, we are also interested in adjusting the local cost (difference of corresponding pitches of reference and test templates) so that perceptually unimportant pitch differences do not affect the DTW optimal path estimate. That is, we would like the path to be biased towards the diagonal transition if the local distances in all directions are comparable. We utilize the same training phrases to decide on local cost bound.

We consider two performances from the dataset (AB and MA) for the training set, to determine the local cost threshold and to learn the constraints. For a given *pakad* class and VQ cluster (category within phrase class), we obtain the DTW paths for all possible pairs of member phrases. To decide on the local cost, the differences between all the corresponding (DTW aligned) pitch values over all the paths are computed. This local cost distribution for each of the *DnDP* and *mnDP* phrase classes is shown in Figure 4.9. We observe that the error is largest near to 0 and falls off quite rapidly beyond 25 cents. We therefore use this value as a local cost lower bound, choosing the diagonal transition whenever the local cost of the diagonal match is within 25 cents of the horizontal and vertical matches for a given location on the path.

For learning the constraints, the DTW paths obtained earlier are recomputed with the local cost bound in place. This means that while computing local cost in DTW, any distance within 25 cents is taken as 0. An outer bound that includes most paths, leaving out only the

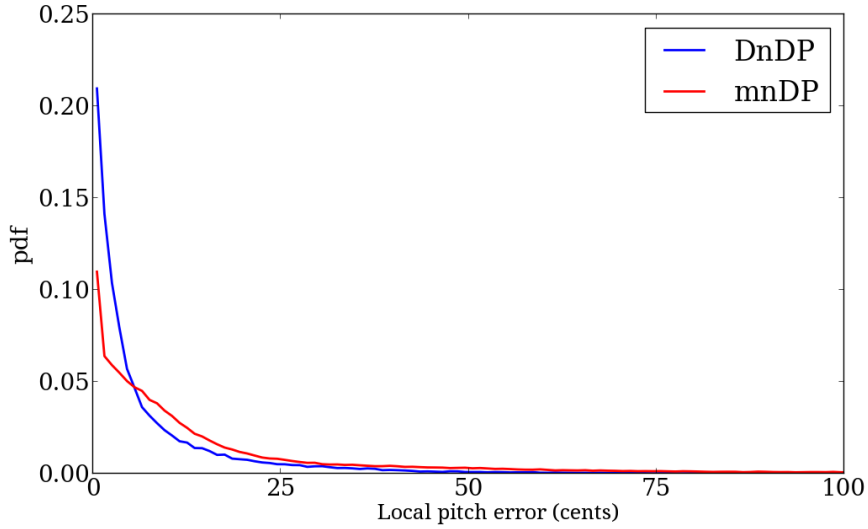
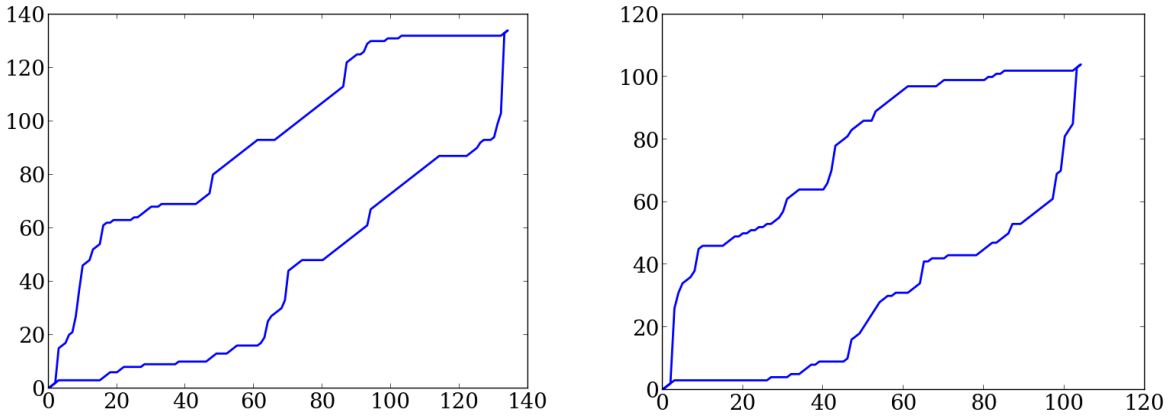


Figure 4.9: Local error distribution between corresponding pitch values after DTW alignment of every pair of phrases within each cluster across *DnDP* and *mnDP* phrase classes from the AB and MA concerts.

obvious outliers, is taken as the global path constraint for the given phrase class and category. Figure 4.10 shows the global path constraint for each of *DnDP* and *mnDP* (first category of each class only). We observe that the global path is wider at the boundaries and displays a narrowing towards the middle in both cases. A closer examination revealed that the narrow region corresponds to the n in *DnDP*, and to the $m \rightarrow n$ glide in *mnDP*. This suggests that the phrases display near constant durations of these sub-segments under overall phrase duration changes. The overall duration changes affect the relatively steady *swara* sub-segments (*D* and *m*) near the phrase boundaries the most. The learned global constraints are compared with standard fixed-width constraints in the experiments.



(a) (b)

Figure 4.10: Learned global constraint obtained by bounding the DTW paths for (a) *DnDP* Cluster 1 with 15% of paths rejected as outliers; (b) *mnDP* Cluster 1with 10% of paths rejected as outliers

4.5 Experiments and Results

We report the results of *pakad* detection experiments carried out as discussed in the previous sections on Hindustani dataset. Evaluation results are presented in terms of retrieval accuracies measured for each of a set of selected phrases on the test data comprised of several phrase classes. We consider the retrieval of *DnDP* and *mnDP* phrases given reference templates obtained by vector quantization on the training set of the same phrase classes from the AB and MA concerts of Table 4.1. The test set includes all the phrases of the corresponding class drawn from the remaining concerts of Table 4.2, plus all phrases not in that class across all the concerts. From Table 4.2, the Hindustani test dataset comprises phrase classes that share the ending P-nyas *swara*. Further, some phrases have several *swara* s in common e.g. *mnDP* and *DnDP*. The test data also includes the *DnDP* segment from Kafi raga (similarly notated but perceptually different), expected to differ in phrase intonation from the raga-characteristic *DnDP* of Alhaiya Bilawal. Thus the design of the Hindustani test data, although small, is challenging. We study the dependence of retrieval accuracy on the choice of pitch quantization and on the DTW global constraint.

The retrieval performance for a given phrase class is measured by applying a threshold to the distribution of DTW distances obtained. Each test phrase contributes a distance value corresponding to the minimum distance achieved between the test pitch curve and the set of

reference templates for the phrase class of interest. In order to increase the evaluation data, 3 rounds of the same test phrases are carried out, each using a different set of reference templates drawn from the training data. The different reference templates are selected from among the training set phrases that lie relatively close to the VQ obtained centroid pitch curves. Figure 4.11 shows such a distribution of the test pitch curve distances across all the reference template sets for the *DnDP* phrase class for chosen reference templates and learned global constraint. Based on the ground-truth labels of the test phrases, two distributions are plotted, viz. one corresponding to positive (true) phrases and the other to negative (i.e. all other phrases including non-raga characteristics). We note that the positive distances are concentrated at low values. The negative distances are widely spread and largely non-overlapping with the positive distances, indicating the effectiveness *DnDP* s of our similarity measure. The negative distribution shows clear modes which have been labeled based on the ground truth of the test phrases. As expected, *GRGP* phrases are most separated while *mnDP* , with more shared *swara* s, is close to the *DnDP* distribution. Finally, the non-phrase *DnDP* overlap most with the positives and are expected to be the cause of false alarms in the raga-characteristic *DnDP* phrase detection.

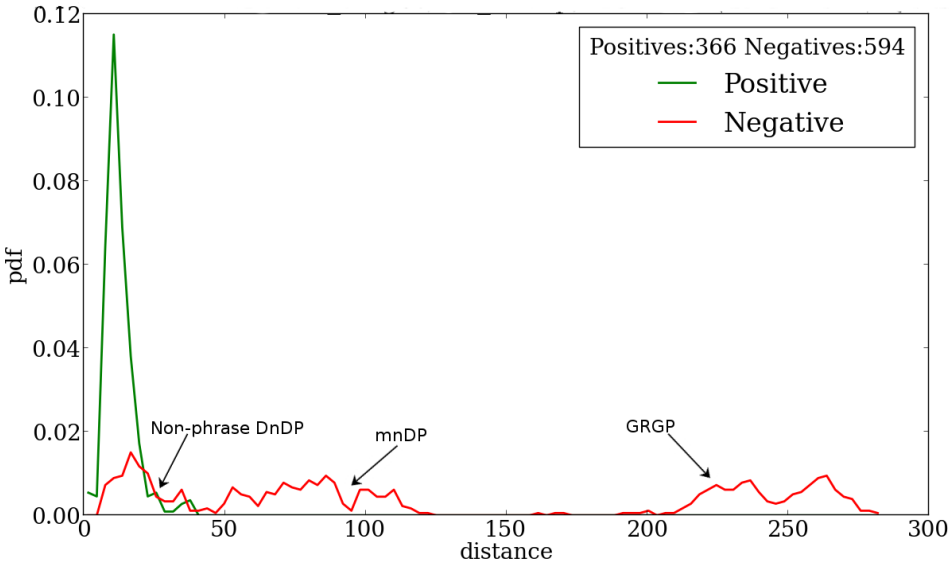


Figure 4.11: Distributions of distances of P -nyas phrases from the *DnDP* raga-characteristic templates. Green: raga-characteristic *DnDP* ; Red: all other phrases

Since an important application of raga-characteristic phrase detection is the retrieval of music based on raga identity, a low false alarm rate would be desirable. Tables 4.3 and 4.4 show the hit rate achieved at a range of acceptable false alarm rates for the *DnDP* and *mnDP* phrases

FA Rate	**Hit Rate**				
	No Constraints	Learned Constraints		Sakoe Chiba	
		15-5%	20-10%	15-5%	20-10%
0.10	0.817	0.822	0.825	0.833	0.828
0.12	0.858	0.882	0.893	0.852	0.872
0.14	0.918	0.921	0.923	0.932	0.934
0.16	0.948	0.948	0.950	0.951	0.945
0.18	0.956	0.956	0.959	0.954	0.951
0.20	0.967	0.967	0.975	0.961	0.969

Table 4.3: Phrase detection accuracies for *DnDP* under various global constraints. (x-y% in the column refers to Cluster-1 constraint rejecting x% and Cluster-2 constraint rejecting y% of the corresponding set of paths).

FA Rate	**Hit Rate**				
	No Constraints	Learned Constraints		Sakoe Chiba	
		20-10%	30-20%	20-10%	30-20%
0.010	0.804	0.979	0.983	0.934	0.965
0.012	0.913	0.988	0.989	0.977	0.982
0.014	0.988	0.993	0.994	0.986	0.990
0.016	0.996	0.995	0.996	0.993	0.994

Table 4.4: Phrase detection accuracies for *mnDP* under various global constraints. (x-y% in the column refers to Cluster-1 constraint rejecting x% and Cluster-2 constraint rejecting y% of the corresponding set of paths).

respectively. The false alarm rates depicted for *DnDP* are higher due the more challenging test dataset for this phrase that includes similarly notated test phrases. We compare the DTW similarity measure obtained with various global constraints, versus that without a constraint. Global constraints obtained by learning as presented in the previous section are compared with the similarly derived Sakoe-Chiba constraint with its constant width (Sakoe and Chiba, 1978). The different global constraints correspond to the fraction of learned paths rejected as outliers based on the maximum distance of a path from the diagonal. Since the clusters were of different sizes, the percentage of paths rejected turned out to be different for the similar extent of deviation from the diagonal. In all cases, the 25 cents error bound was applied in order to bias the DTW towards the diagonal direction.

FA Rate	**Hit Rate**		
	Unquantized	q12	q24
0.10	0.825	0.863	0.849
0.12	0.893	0.888	0.896
0.14	0.923	0.904	0.921
0.16	0.950	0.923	0.939
0.18	0.959	0.937	0.951
0.20	0.975	0.956	0.962

Table 4.5: Phrase detection accuracies for various quantized representations for *DnDP*. The learned constraint (20-10%) is used here.

FA Rate	**Hit Rate**		
	Unquantized	q12	q24
0.010	0.983	0.976	0.959
0.012	0.989	0.988	0.977
0.014	0.994	0.990	0.987
0.016	0.996	0.994	0.994

Table 4.6: Phrase detection accuracies for various quantized representations for mnDP. The learned constraint (30-20%) is used here.

We observe that retrieval performance of DTW with global constraints clearly surpasses that without constraint. The learned constraints provide for superior or similar hit-rate across all FA rates (except one in case of *DnDP*). Tables 4.3 and 4.4 use unquantized pitch for the melodic representation. Tables 4.5 and 4.6 compare the performances with various pitch quantization at the fixed best-performing selected global constraint. We observe that unquantized pitch works best for most conditions.

4.6 Summary

In this work on *pakad* detection, we addressed the classification of segmented phrases. Given the lack of a basis for note segmentation, the continuous pitch curve obtained by pre-dominant pitch detection of vocal music audio was treated as a time series. Similarity modeling with DTW based distance showed promising results on our datasets. The observed variability in within-class phrase intonation is captured by supervised learning of the DTW global constraints from training data. Considering that the flexibility available to the artiste in rendering a raga-characteristic phrase is constrained partly by the need to retain its distinctiveness with respect to other ragas, it would be worthwhile to learn the melodic representation by training on a larger corpus of the same raga including an anti-corpus of different ragas where the similar phrase occurs.

Further work should incorporate volume and timbre dynamics so that the exploration of the constancy of raga-characteristic phrases. Tempo is known be a strong determiner of melodic shape, and the study of phrase intonation at high tempos is needed. Moving on from the time-series representation, a more event-based representation of the phrase in terms of basic melodic shapes is likely to be less affected by the allowed improvisational changes in phrase intonation. The potential of including more explicit music knowledge in motif detection will be explored in future. Finally, the important sub-task of melodic segmentation must be considered. Cues to melodic phrase ending often, but not always, include the occurrence of a final resting note. The modeling of the perception of closure associated with phrase ending is a hitherto unexplored problem.

Chapter 5

Raga Similarity Detection from Composition Notation

Raga being one of the most prominent categorization aspect of Hindustani music, detecting similarities between ragas can be useful to many Hindustani music specific tasks like music search, music recommendation, automatic analysis of large-scale musical content etc. Generally similarity between ragas is inferred through attributes associated with the ragas. For instance, in Hindustani music, one classification of ragas is based on the tonal material involved and termed as *thaat*. There are 10 *thaats* in Hindustani music (Chakraborty et al., 2014). Emotion, time of day, *jati*, *vadi*, *samvadi* etc. are the other important attributes. Most of the accepted similarities between ragas encompass the similarities in many of these attributes. But these similarities cannot always be derived exclusively from these attributes. Quantifying raga similarity becomes extremely challenging as it demands assimilation of both *intrinsic* (*viz.*, notes, tempo) and extrinsic (*viz.* raga singing-time, emotions conveyed) properties of ragas. In other words, the similarities between ragas is not always a function of their attribute-wise similarities.

We investigate on extracting raga similarity from two distinct data sources, viz. composition (*bandish*) notations and textual discussions on Hindustani ragas. Both the methods learn representation for each raga, and the similarity between the representations is taken as the similarity between the ragas. This chapter discusses the approach with melodic content available as *bandish* notation and the latter is discussed in the next chapter. We do not intent to extract the same similarity information from both the data sources; but similarities from two different perspectives. For the same reason, we do not compare the results of both. Even though there are many overlaps, the similarities captured from discussions are considered different from simi-

larities captured from melodic content. Most of the similarities a trained musician can perceive may not always be identified from melodic content. Many of the available online discussions and books will have an account on it. But text content does not serve to extract similarities for which melodic attributes are the prime.

In Hindustani music, notation provides an abstract framework for a composition performance. The approach designed for extracting raga similarity from *bandish* notation directly relies on melodic attributes available in the notation. Based on the hypothesis that notes in a particular raga are characterized by the company they keep, we design and train several deep recursive neural network variants with Long Short-term Memory (LSTM) units to learn distributed representations of notes in ragas from *bandish* notations. We refer to these distributed representations as note-embeddings. Note-embeddings, as we observe, capture a raga's identity, and thus the similarity between note-embeddings signifies the similarity between the ragas.

This chapter is organized as follows. The discussion begins with the motivation and central idea in Section 5.1. The neural network architecture for learning note-embeddings and extracting similarities from these representation are discussed in Sections 5.2 and 5.3. The baselines to compare with our approach are described in Section 5.4. The *bandish* notation dataset is described in Section 5.5, followed by pre-processing of data in Section 5.6. Sections 5.7 and 5.8 discuss the experiments and results.

5.1 Raga Similarity Based on Notation: Motivation and Central Idea

A raga-similarity method solely based on notational (melodic) information can be quite relevant to computational music tasks involving Indian classical music. While the general notion of raga similarity is based on various dimensions of ragas like *thaat*, *prahar*, *jati*, *vadi*, *samvadi* etc., the similarities perceived by humans (musicians and expert listeners) is predominantly based upon the melodic structure.

Theoretically, the identity of a raga lies in how certain notes and note sequences (called phrases) are used in its compositions. We hypothesize that capturing the semantic association between different notes appearing in the composition can possibly reveal the identity of a raga. Moreover, it can also provide insights into how similar or dissimilar two ragas can be, based on how similar or dissimilar the semantic associations of notes in the compositions are. We

believe, notes for a specific raga can be represented in distributed forms (such as vectors), reflecting their semantic association with other notes in the same raga (analogous to words having distributed representations in the domain of computational linguistics (Mikolov et al., 2013a)). These representations could account for how notes are preceded and succeeded by other notes in compositions.

Formally, in a composition, a note $x \in V$ (where V represents a vocabulary all notes in three octaves) can be represented as a d dimensional vector that captures semantic-information specific to the raga that the compositions belong to. Such distributed note-representations, referred to as note-embeddings can be expected to capture more information than other forms of sparse representations (like presenting notes with unique integers). A raga can, thus, be characterized by a $|V| \times d$ embedding matrix in which each row represents a d dimensional note-embedding for a single note. We propose a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) based architecture that is motivated by the the work of Huang and Wu (2016) to learn note-embeddings characterizing a particular style of music. We learn note-embeddings for each raga separately from the compositions available for the raga.

How can note-embeddings help capture similarities between ragas? We hypothesize that embeddings learned for a given note for similar ragas will have more similarity. For example, the representation for note *Ma-elevated* (equivalent note *F#* in C-scale) in raga Yaman can be expected to be very similar to that of Yaman Kalyan as both of these ragas share very similar melodic characteristics.

5.2 Neural Network Architecture for Learning *Note-Embeddings*

We design a deep recurrent neural network (RNN), with bi-directional LSTMs as recurrent units, that learns to predict the forth-coming notes that are highly likely to appear in a *bandish* composition, given input sequences of notes. This is analogous to neural language models built for speech and text synthesis (Mikolov et al., 2011). While our network tries to achieve this objective, it learns distributed note representations by regularly updating the note-embedding matrix. The choice of this architecture is due to the facts that (a) for sequence learning problems like ours, RNNs with LSTM blocks have proven useful (Sutskever et al., 2014; Eck and Schmidhuber, 2002), and (b) in Hindustani music a note rendered at a moment has dependence on patterns preceding and succeeding it, motivating us to use bi-directional LSTM.

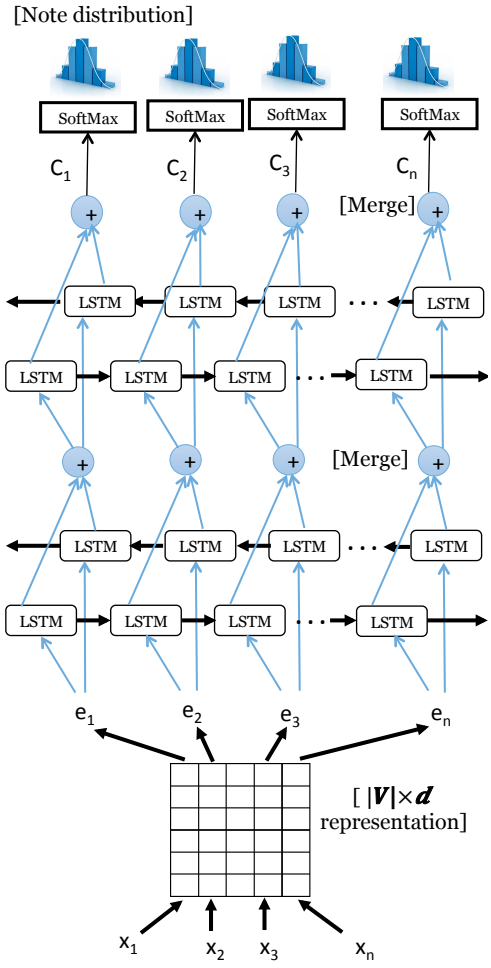


Figure 5.1: Bi-directional LSTM architecture for learning note-embeddings

The model architecture is shown in Figure 5.1. Supposing that a sequence in a composition has n notes (n to be kept constant by padding wherever necessary), denoted as $x_1, x_2, x_3, ..., x_n$, where $\forall i \in n, x_i \in V$. The note x_i can be represented in one-hot format, with the j^{th} component of a $|V|$ dimensional zero-vector set to 1, if x_i is the j^{th} element of vocabulary V. Each note is input to a note-embedding layer W of dimension $|V| \times d$ where d is the note-embedding dimension. W is initialized randomly and is updated via back-propagation during training. The output of this layer is a sequence of embeddings e_i of dimension d, obtained by performing a matrix multiplication between x_i with W. The embedding sequences $e_1, e_2, e_3, ..., e_n$ are input to two layers of bi-directional LSTMs.

For each time-step ($i \in n$), the context-representations learned by the outer-bidirectional LSTM layer (C_i) is passed through a `softmax` layer that computes the conditional probability distribution of all possible notes given the context representations given by LSTM layers.

For each time-step, the prediction of the forthcoming note in the sequence is done by choosing the note that maximizes the likelihood given the context i.e.

$$\hat{x} = \underset{j \in |V|}{argmax} \quad P(x_{i+1} = v_j | C_i) \tag{5.1}$$

where C_i is the merged context representations learned by the forward and backward sequences in the bi-directional LSTM layers. Probability of a note at a time-step is computed by the `softmax` function as,

$$P(x_{i+1} = v_j | C_i) = \frac{exp({U_j}^T C_i + b_j)}{\sum_{k=1}^{|V|} exp({U_k}^T C_i + b_k)} \tag{5.2}$$

where U is the weight matrix in the `softmax` layer and b_j is bias term corresponding to note v_j.

The embedding layer is initialized randomly and during training, errors (in terms of cross-entropy) are back propagated upto the embedding layer, resulting in the updation of the embedding-matrix. Cross-entropy is computed as,

$$\frac{1}{M \times T} \sum_{i=1}^{M} \sum_{t=1}^{T} cross\_entropy(y_t^i, \hat{y}_t^i) \tag{5.3}$$

$$cross\_entropy(y, \hat{y}) = -\sum_{p=1}^{|V|} y_p \; log \; \hat{y}_p \tag{5.4}$$

Where M is the number of note sequences in a raga and T is the sequence length. y_t^i denotes the expected distribution of i^{th} note sequence at time-step t and $\hat{y}_t^i$ denotes the predicted distribution.

5.3 Raga Similarities from Note-Embeddings

For each raga our network learns a $|V| \times d$ matrix representing $|V|$ note-embeddings. We compute (dis)similarity between two ragas by computing pairwise cosine distance between embedding vectors of every note in V and then averaging over all notes. This is based on the assumption that distributed representations of notes (as captured by the embeddings) will be similar across ragas that are similar. The choice of cosine similarity (or cosine distance) for computing the similarity between the note-embeddings is driven by its robustness as a measure of vector similarity for vectors and its predominant usage for measuring word embedding similarity (Mikolov et al., 2013c). Appropriate distance measures have been adopted for non-LSTM based baselines.

5.4 Baselines for Comparison

We compare our approach with a few baseline approaches to confirm the validity of our approach.

5.4.1 N-gram based approach

The N-gram based baseline creates an n-gram profile based on the count of each n-gram from the available compositions in a raga. We compute the n-gram for n ranging from 1 to 4. The distance between two ragas is computed using the out-of-place measure described in Cavnar et al. (1994). Out-of-place measure depends on the rank order statistics of the two profiles. It computes how far 2 profiles are out-of-place *w.r.t* the n-gram rank order statistics. The distance is taken as the $L2 - norm$ of all the n-gram rank differences, normalized by the number of n-grams. Intuitively, the more similar two ragas are, more would the N-gram profiles overlap, reducing the $L2 - norm$.

5.4.2 Pitch Class Distribution (PCD)

This method computes the distribution of notes from the count of notes in a raga's *bandish* dataset. 36 notes (across 3 octaves) are considered separately for computing PCD. As the method describes, sequence information is not captured here. The similarity distance between two ragas is computed by taking the euclidean distance between the corresponding pitch class

distributions; the assumption is that each pitch class two similar ragas will share similar probability value, thereby reducing the euclidean distance. For the raga recognition task by Chordia (2006), euclidean distance is used for computing the distance between pitch class distributions in one of their approaches. This baseline is to verify the relevance of sequence information in capturing raga similarity.

5.4.3 Uni-directional LSTM

The effectiveness of a bi-directional LSTM for modeling Hindustani music is verified with this baseline. The architecture is same as described in Figure 5.1, except for the replacement of bi-directional LSTMs with uni-directional LSTMs. Since there is only forward pass in uni-directional LSTM, the merge operation in bi-directional LSTM design is not required here.

5.5 Dataset

Our experiments are carried out with the Hindustani *bandish* dataset available from `swarganga.org`, created by Swarganga music foundation. This website is intended to support beginners in Hindustani music. This has a large collection of Hindustani *bandishes*, with lyrics, notation, audio and information on raga, *tala* and *laya*. Figure 5.2 shows a *bandish* instance from swarganga. The name of this *bandish* is *'jaane naa jaane haree'* in raga Adana and in *tintal* (16 beats cycle). The first row contains the *bol* information which details the tabla strokes corresponding to the *tala* of the *bandish*. Other rows have lyrics (bottom) along with the notes (top) corresponding to the lyrical sections. Each row corresponds to a *tala* cycle. A note followed by a single quotation at the right shows it is in the higher octave and a single quotation at the left implies lower octave. Notes mentioned within parenthesis are *kan* notes (grace notes). Each column represents a beat duration.

#bandishes	**#ragas**	**#notes**	**#kan swaras** (grace notes)
2955	144	2,95,411	50,749

Table 5.1: Dataset

1 dhA +	2 dhiM	3 dhiM	4 dhA	5 dhA *2*	6 dhiM	7 dhiM	8 dhA	9 dhA *o*	10 tiM	11 tiM	12 tA	13 tA *3*	14 dhiM	15 dhiM	16 dhA
						R' jaa	S' ne	R' naa	n jaa	S' ne	n ha	P ree	m bha	P ja	S' na
S' bee	R'S' -	d -	d -	n -	P na	n -	mP -	m naa	- -	P da	n shru	g tee	gm ma	R dhu	S ra
m su	- -	P ra	d la	n ya	S'R' lee	g'R' -	S' na								
								m ra	P sa	d raM	d ga	n ai	n so	S' sa	S' maa
n Chaa	S' yo	R' aa	S' sa	d maa	d -	n -	P na	S'm' ha	g'm' ree	R' ha	S' ree	d su	d mi	n ra	P na
m daM	P ga	d sa	n ba	S'R' ra	g'R' see	S'n ya	S' na								

Figure 5.2: A *bandish* instance from swarganga website.

From this dataset we have considered 144 ragas for our study which are represented well with sufficient number of *bandishes*. Table 7.1 presents dataset statistics.

5.6 Data Pre-processing

We take all *bandishes* in a raga for training the note-embeddings for the raga. *Kan* notes are also treated in the same way as other notes in the composition, since the *kan* notes also follow the raga rules. The notes are encoded into 36 unique numbers. The notes corresponding to a *tala* (rhythm) cycle is taken as a sequence. The input sequence length is determined by taking the average length of the sequences in a raga dataset; zero-padding (to the left) and left-trimming of sequences are applied to sequences shorter and longer than the average length respectively. If the length of a sequence is more than double the defined sequence length, it is split into 2 separate sequences.

5.7 Experiments

We now share our experimental details.

5.7.1 Evaluation Methods

We rely on 2 different evaluation methods to validate our approach. The first one is based on perplexity that evaluates how well a note-sequence generator model (neural-network based, n-gram based etc.) can predict a new sequences in a raga. Since note-embeddings are an integral part of our architecture, a low-perplexed note-sequence generator model should learn more accurate note embeddings. The second method relies on clustering of ragas based on different raga-similarity measures computed using our approach and baselines.

Perplexity: Perplexity for a language model (Bahl et al., 1983), is computed based on the probability values a learned model assigns to a validation set (Clarkson and Robinson, 2001). For a given model, perplexity (PP) of a validation set with notes $N_1, N_2, ..., N_n$ is defined as

$$PP(N_1, N_2, ..., N_n) = \sqrt[n]{\frac{1}{P(N_1, N_2, ..., N_n)}} \tag{5.5}$$

where $P(N_1, N_2, ..., N_n)$ is the joint probability of notes in the validation set. A better performing model will have a lower perplexity over the validation set. For each raga dataset, perplexity is measured with a validation set ($\sim$100 notes) taken from the dataset. For the LSTM based methods, the learned neural model provides the likelihood of a note, whereas the n-gram baseline uses the learned probabilities for different n-grams.

Thaat	Ragas
Kalyan	Shuddha Kalyan, Yaman Kalyan, Yaman
Marwa	Marwa, Puriya, Sohni
Bilawal	Alhaiya Bilawal, Bihag, Shankara
Kafi	Kafi, Bageshree, Bhimpalasi
Bhairav	Bhairav
Asavari	Jaunpuri

Table 5.2: Thaat based grouping of the selected ragas

Clustering: For this evaluation, we take 14 ragas for which similarities between all the ragas and subsets of these ragas are known. These similarities are determined with the help of a professional Hindustani musician. The selected ragas are Shuddha Kalyan, Yaman Kalyan, Yaman, Marwa, Puriya, Sohni, Alhaiya Bilawal, Bihag, Shankara, Kafi, Bageshree, Bhimpalasi, Bhairav and Jaunpuri. The first clustering (`Clustering 1`) checks if all the 14 ragas are getting clustered according to their *thaat*. *Thaat* wise grouping of these 14 ragas are shown in Table 5.2. Since there are 6 different *thaats*, k is taken as 6 for this clustering. For the other clusterings, different subsets of ragas are selected according to the similarities to be verified. Other similarities and the ragas chosen (from the 14 ragas) to verify that are as listed below.

- `Clustering 2`: Sohni is more similar to Yaman and Yaman Kalyan compared to ragas in other *thaats* because they share the same characteristic phrase (*MDNS*). To verify this, Sohni, Yaman, Yaman Kalyan, Kafi, Bhairav are considered taking k=3 and we expect the first 3 ragas to get clustered together and, Kafi and Bhairav in 2 different clusters.

- `Clustering 3`: Within Kafi *thaat*, Bhimpalasi and Bageshree are more similar compared to their similarity with Kafi because of the similarity in these ragas' characteristic phrases (*mDnS, mPnS*). To verify this, these 3 ragas are considered for clustering taking k=2 and we expect Bhimpalasi and Bageshree to get clustered together and Kafi in another cluster.

- `Clustering 4`: Raga Jaunpuri is more similar to Kafi *thaat* ragas because they differ only by a note. To verify this, Jaunpuri, Kafi, Bageshree, Bhimpalasi, Bhairav, Shuddha Kalyan, Puriya, Bihag are considered taking k=5. We expect Jaunpuri to be clustered

together with Kafi, Bageshree and Bhimpalasi, and the other ragas in 4 different clusters.

We apply these four clustering methods on our test dataset and evaluation scores pertaining to each clustering method is averaged to get a single evaluation score.

5.7.2 Setup

For the experiments, we consider notes from 3 octaves, amounting to a vocabulary size of 37 (including the null note). The common hyper-parameters for the LSTM based methods (our approach and one of the baselines) are kept the same. The number of LSTM blocks used in the LSTM layer is set to the sequence length. Each LSTM block has 24 hidden units, mapping the output to 24 dimensions. For all our experiments, embedding dimension is empirically set to 36. We use `tensorflow` (version: 0.10.0) (Abadi et al., 2016) for the LSTM implementations.

While training the network, the perplexity of the validation set is computed during each epoch and used for setting the early-stopping criterion. Training stops on achieving minimum perplexity and the note-embeddings at that instance are taken for our experiments.

For the clustering baseline, we employ one of the hierarchical clustering methods, agglomerative clustering (*linkage:complete*). In our setting, a hierarchical method is preferred over K-means because, K-means work well only with isotropic clusters (Nagy, 1968) and it is empirically observed that our clusters are not always isotropic. Also when experimented, the clustering scores with K-means are less compared to agglomerative clustering for all the approaches. For implementing the clustering methods we use scikit-learn toolkit (Pedregosa et al., 2011).

5.8 Results

Before reporting our qualitative and quantitative results, to get a feel of how well note-embeddings capture raga similarities, we first visualize the 37×36 note-embedding matrices by plotting their heatmaps, higher intensity indicating higher magnitude of the vector component. Figure 5.3 shows heatmaps of embedding matrices for three ragas *viz.* Yaman Kalyan, Yaman and Pilu. Yaman Kalyan and Yaman are more similar to each other than Pilu. This is quite evident from the embedding heatmaps which appear similar for the first two, and different for the third one.

The results of quantitative evaluation is now reported with the evaluation methods described in Section 5.7.1. Further, a manual evaluation is done with the help of trained Hin-

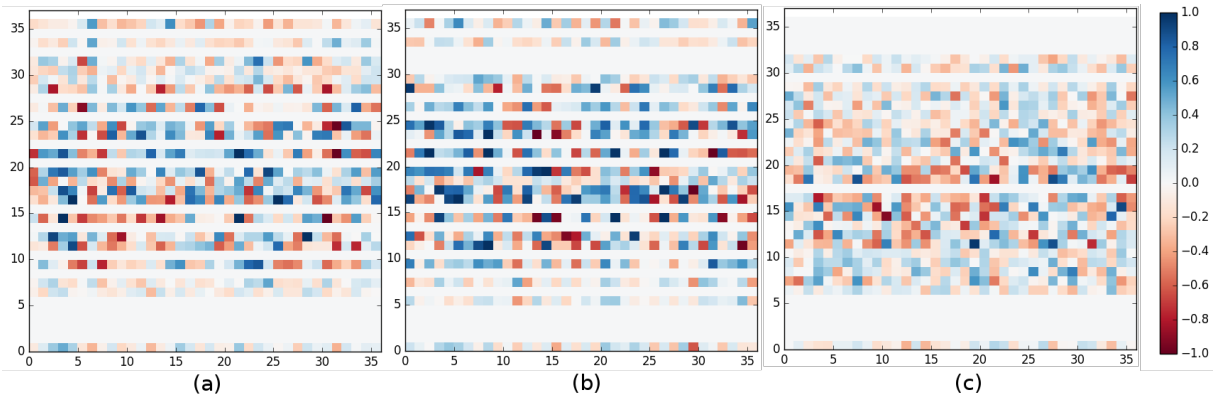


Figure 5.3: Note-embeddings visualization of (a) Yaman Kalyan (b) Yaman (c) Pilu

dustani musician considering all the 144 ragas mentioned in the dataset, to better understand the distinctions between bi-LSTM and uni-LSTM. Table 5.3 shows perplexity values (averaged

Experiment	**Perplexity**
N-gram	6.39
uni-LSTM	6.40
bi-LSTM	**2.31**

Table 5.3: Results: Comparison with perplexity on validation set (Best performance in bold)

across all the ragas in the dataset) with the validation set for our approach (bi-LSTM) and the baseline approaches with n-gram and uni-directional LSTM (uni-LSTM). Lower the perplexity, better the performance. We can not report perplexity for the PCD approach as the likelihood of the notes (and hence, the perplexity of the model) can not be determined with PCD. We observe that the perplexity values of n-gram and uni-LSTM are quite similar. The lower perplexity value with bi-LSTM shows its capability in generating a new notes sequence adhering to the raga rules. This shows the performance advantage of bi-LSTM over the baselines on note-sequence generation task, thereby providing indications on the goodness of the note-embeddings learned. Moreover, the bi-LSTM model, having the lowest perplexity, is able to capture the semantic association between notes more accurately, yielding more accurate note-embeddings.

Table 5.4 shows the results of clustering using a standard set of metrics for clustering, *viz.* homogeneity, completeness and V-measure (Rosenberg and Hirschberg, 2007). The clustering scores with n-gram and PCD baselines show their inability towards identifying the known similarities between the ragas. The bi-LSTM approach performs better compared to the baselines;

Experiment	Homogeneity	Completeness	V-measure
N-gram	0.3973	0.4036	0.4004
PCD	0.6430	0.6488	0.6451
uni-LSTM	0.7828	0.7858	0.7843
bi-LSTM	**0.9008**	**0.9069**	**0.9038**

Table 5.4: Results: Comparison of clustering results with different clustering metrics (Best performance in bold)

the performance of uni-LSTM baseline is comparable with bi-LSTM approach. On analyzing each individual clustering, we observed,

- N-gram approach does not do well for all the individual clusterings, resulting in poor clustering scores compared to other approaches. A relatively better performance is observed only with `Clustering 4`.

- PCD has better scores compared to n-gram as it out-performs n-gram with a huge margin in `Clustering 1`. PCD's performance in `Clustering 1` is superior to the LSTM approaches as well. However, its performance is quite inferior to that of other approaches in the other three clustering settings. PCD's ability in modeling notes distribution efficiently helps in *thaat* based clustering (`Clustering 1`), because *thaat* based classification quite depends on the distribution of tonal material.

- uni-LSTM performance is better than bi-LSTM in `Clustering 1` where the ragas are supposed to be clustered according to the *thaat*. But it fails to cluster Sohni, Yaman and Yaman Kalyan in the same cluster, leading to poor performance in `Clustering 2`.

- Even though bi-LSTM gives slightly lower scores with `Clustering 1`, it does perfect clustering for the other three clustering schemes. This gives an indication on the capability of bi-LSTM approach for identifying melodic similarities beyond *thaat*.

Overall, these observations show the practicality of both the LSTM based methods to learn note-embeddings with the aim of identifying raga similarity.

Figures 5.4 show Multi-Dimensional Scaling (MDS) (Borg and Groenen, 2003) visualizations showing the similarity between note-embeddings of the selected 14 ragas (same color

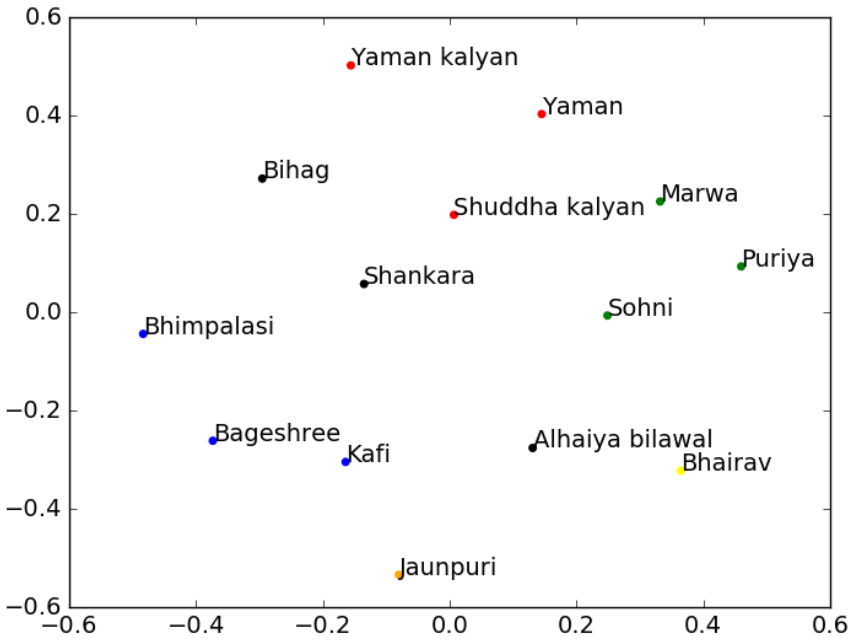


Figure 5.4: MDS visualization of bi-LSTM note-embeddings similarities

specifies same *thaat*) with bi-LSTM approach. These visualizations give an overall idea on how well the similarities are captured. The finer similarities observed in the clustering evaluations are not clearly perceivable from these visualizations.

We have also carried out separate experiments by including note duration information along with the notes by pre-processing the data, but the performance is worse compared to the reported results. Chordia (2006) has also reported that weighting by duration had no impact on their raga recognition task.

To confirm the validity of our approach, one expert musician checked the MDS visualizations of similarities between all 144 ragas with bi-LSTM and uni-LSTM approaches. The musician identified clusters of similar ragas in both the visualizations matching with his musical notion. A few observations made are: Asavari *thaat* ragas appear to be closer to each other with bi-LSTM compared to uni-LSTM. Also Miyan ki todi, Multani, Gujari Todi which are very similar ragas are found closer in bi-LSTM. But the same *thaat* ragas Marwa, Puriya and Sohni are found to be more similar to each other with uni-LSTM.

5.9 Summary

In this chapter we discussed our approach to detect similarities between ragas from *bandish* notation. The discussed approach learns a representation for each raga in the form of note-

embeddings. The learned representation capture the identity of a raga and the similarities between the representations give the similarities between the corresponding ragas. The results shows the effectiveness of the methods in revealing the similarities between ragas. Perplexity based evaluation shows the superior performance of bi-directional LSTM method over unidirectional-LSTM and other baselines. The clustering based evaluation also confirms this, but it also shows that the performance of uni-directional approach is comparable to the bi-directional approach for certain cases.

As future work, we plan to design a network to handle duration information effectively for better learning of note-embeddings. The current experiments take one line in the *bandish* as a sequence. We plan to experiment with more meaningful segmentation schemes like lyrical phrase delimited by a long pause. The utility of this approach is not confined only to raga similarity; it can also be extended to verify if a given *bandish* complies with the raga rules. This immensely benefits to Hindustani music pedagogy; for instance, it helps to select the right *bandish* for a learner.

Chapter 6

Raga Similarity Detection from Textual Discussions

This chapter continues the discussion presented in the last chapter with an approach to extract raga similarities from text discussions. As mentioned, many of the perceived raga similarities by musicians are beyond the raga attributes and the similarities directly extractable from melodic content. Most of these perceived similarities are expressed in discussions available as books and online articles. The proposed approach here, takes textual discussion as data source and utilizes distributional hypothesis to learn representation for a raga word. When the former approach learns representation for the notes in a raga from the notes context, this approach learns representation for a raga word in the form of word vector from the context information in the text.

Many text resources with description on ragas available in the form of books, websites, forums, etc (Parrikar, 2000; Swarganga, 2004; van der Meer, 2012) have excellent depictions of raga similarities. The textual discussions based approach targets extraction of similarities which are generally accepted by musicians and likely to appear in text discussions. Ragas perceived as similar can be found closer in text discussions, and also other common words which occur along with raga names can be same for similar ragas. Here we learn word vectors for the raga words from textual descriptions and discussions on Hindustani ragas. Word vectors are continuous vector representations of words, learned with very large datasets capturing the similarity between the words (Mikolov et al., 2013a). In this approach, we propose a modification to Mikolov's word vector learning method, to learn raga word vectors efficiently from scarce resources. The method learns word vectors of raga words (absent in the general pool of word

vectors) with the help of this specific dataset and pre-trained word vectors trained on general English corpus like Google word vectors.

This chapter is organized as follows. Section 6.1 describes the concept of distributional semantics which forms the foundation of Mikolov's approach to word vector learning. Section 6.2 discusses Mikolov's approach and our modification to this approach for learning domain-specific word vectors. Section 6.3 describes the datasets for the experiments. The experiments and results are discussed in Section 6.4 and Section 6.5 respectively. The applicability of this approach for any domain is discussed in Section 6.6.

6.1 Distributional Semantics

Distributional semantics aim at learning semantic representations for words from the meaning of the words. This representation depending on the context-based representation is relevant to NLP tasks for which probability analysis of linguistic distributions is valid (Lenci, 2008). Here the combinatorial behaviour of a word decides its meaning or in other words, words in similar contexts have similar meanings (Goldberg and Levy, 2014). This learning is dependent on the word usage statistics got from a very large corpus. Any approach to learn word semantic representation take the co-occurrence matrix from the text content and transform them to better representation after performing required normalization and dimensionality reduction. There has been a few attempts to learn semantic representation for words (Pennington et al., 2014; Eckart and Young, 1936; Deerwester et al., 1990; Bullinaria and Levy, 2007), but the introduction of neural network based approaches (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov et al., 2013b) brought in revolutionary changes in the use of word representations for NLP tasks. Mikolov's method realized through the `word2vec` program is widely used because of its efficiency and the significant performance improvement it brought to many NLP tasks (Levy et al., 2015).

6.2 Neural Network Architecture

6.2.1 Mikolov's Architecture

In order to capture the semantic similarity between the words, the word vectors learning approach by Mikolov et al. computes continuous vector representation of words from large cor-

pus. The network architecture is designed to predict the word/words in a sequence given the context words or a word in the sequence. The continuous bag-of-word model predicts the target given the context words, whereas the skip-gram model predicts the context words given the target word. (Mikolov et al., 2013a; Rong, 2014)

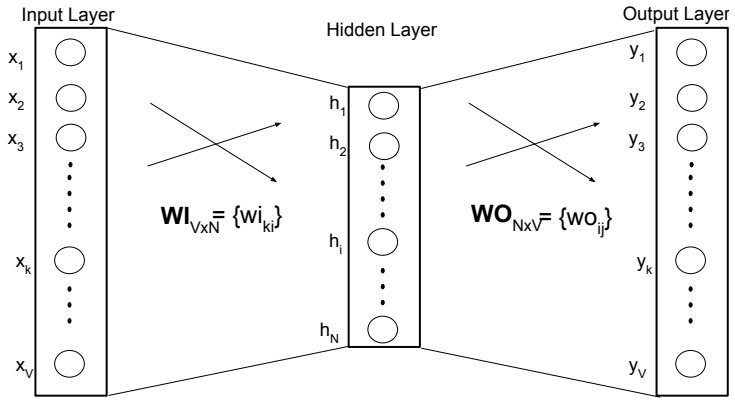


Figure 6.1: Mikolov's approach neural network model (Rong, 2014)

Figure 6.1 shows Mikolov's architecture. This is a feed-forward neural network, with input layer, one hidden layer and output layer. The size of the input layer is V (size of vocabulary) and hidden layer size is N (word vector dimension). The weights between input and hidden layer is represented as $WI \in \mathbb{R}^{V \times N}$ and weights between hidden layer and output layer is represented as $WO \in \mathbb{R}^{N \times V}$. For word vector training in Mikolov's approach, the weight matrices are randomly initialized and WI is taken as the word embeddings for the V words post training.

6.2.2 Proposed Approach

For generating raga names word vectors, we utilize textual resources having discussions on Hindustani ragas and available pre-trained word vectors like Google word vectors. Here, re-trained word vectors employed for this is referred to as base word vectors. Google word vectors are 300 dimensional vectors trained with 3 million words and phrases. The size of available text discussing ragas is not sufficient enough to learn word vectors with the original `word2vec` approach. To address this, our approach introduces a modification to the training of neural network architecture by Mikolov.

In the proposed method, WI and WO are pre-loaded with base word vectors for the words available in the base word vectors. For the rest of the words including raga names, corresponding WI and WO vectors are randomized as done in Mikolov's approach. A list

of raga names are made available and while training the word vectors, only the WI and WO vectors corresponding to the raga names are updated.

The word vectors learned for the common words from a larger context bear the expected meaning of the word which do not differ much in this specific domain text. But word vectors for common words trained from a small dataset may not encompass the actual meaning and will eventually affect the training of dataset specific rare words. The proposed modification to word vector training helps to prevent the common word vectors from getting changed and eventually affecting the raga names word vectors.

6.3 Datasets

Table 6.1 shows the details of the datasets. The datasets are used in combination for different experiments. *Swaraganga* (Swarganga, 2004) and *Parrikar* Parrikar (2000) has raga related information extracted from 2 different websites on Hindustani music. *Wikipedia* dataset has Wikipedia pages on Hindustani ragas and *Hindustani Music In the 20th Century* is a book by Wim van der Meer (van der Meer, 2012). *Swarganga* and *Wikipedia* have description of the ragas in a more formal way explaining the main attributes and some relations, whereas *Parrikar* and *Hindustani Music In the 20th Century* have more subjective description from musicologist's point of view. The first 2 datasets being small in size are used in combination and the rest 2 of significant size are added to this in the other experiments. 154 Hindustani ragas are considered for this study.

Dataset	**Size (#words)**
Swarganga website (*sganga*)	17912
Wikipedia (*wiki*)	24094
Hindustani Music In the 20th Century (*meer*)	99252
Parrikar website (*parrikar*)	115796

Table 6.1: Details of Datasets

6.4 Experiments

As discussed, our method updates word vectors for specific words in a dataset, preserving common words present in pre-trained word embeddings. In this work, the dataset-specific words are the raga words. `word2vec`[1] (as per Mikolov's approach) released by Google (Goldberg and Levy, 2014) is modified to accommodate the proposed design. We evaluate the performance of this method based on the quality of similarity between ragas observed through raga word vectors. The *cosine similarity* between word vectors is taken as the measure of similarity. Many of the tasks involving word vectors have identified cosine similarity as the effective similarity measure (Mikolov et al., 2013c). Experiments are performed with combinations of datasets mentioned in Table 6.1. For all our experiments with Mikolov's approach (baseline) and our approach, parameters are set to default values[2] except for *min-count*, which defines the minimum frequency required for a word to be considered. *min-count* is set to less than 4 for the experiments involving raga datasets to ensure the inclusion of meagerly occurring raga names and related words to the vocabulary. All the experiments are performed only with word embedding dimension of 300. This is constrained by the embedding dimension of the pre-trained word vectors from Google. Since we preload the Google word vectors which are available only in dimension 300, the word vectors learned by our incremental approach also have to be 300. To make the experiments comparable, all the other experiments which are not dependent on Google word vectors also have to follow the same dimension.

Two baseline methods are devised with Mikolov's approach. First baseline (**BL-1**) generates word vectors from raga datasets alone, whereas the second baseline (**BL-2**) combines general English corpus with raga datasets. The general English corpus is formed combining DBPedia short abstracts and long abstracts, Europarl English monolingual corpus (Koehn, 2005), news commentary corpus (released for WMT 2011 shared task). Table 6.2 shows the results of the baseline experiments.

6.4.1 Quantitative Evaluation

The non-availability of ground truth for this task makes the evaluation challenging. We designed an evaluation method considering one strong attribute of similarity between ragas, the *thaat*.

[1] http://word2vec.googlecode.com/svn/trunk/

[2] word2vec parameters: window=10, iter=3 and negative=10

This evaluation method is based on the assumption that majority of the ragas identified similar to a particular raga will belong to the same *thaat* (Rao and Rao, 2014b). The evaluation method described in Algorithm 1, takes 10 most similar ragas to each raga and computes the score based on the number of similar ragas belonging to the same *thaat* as that of input raga (r). For our experiments this serves as suitable metric for comparison.

Algorithm 1 Compute evaluation score

Require: list of raga names ($ragas$), word vectors

Ensure: $score$

1: $score \leftarrow 0$

2: $total\_no \leftarrow 0$

3: **for all** r in $ragas$ **do**

4: $sim\_ragas \leftarrow$ 10 most similar ragas to r based on cosine similarity between raga word vectors

5: $t \leftarrow$ thaat of r

6: $score+ = \sum_{r_s \in sim\_ragas} belongs\_to(r_s,\ t)$

7: $total\_no+ =$ length(sim_ragas)

8: $score \leftarrow score/total\_no$

6.4.2 Qualitative Evaluation

Qualitative evaluation is also done with the help of a trained Hindustani musician, to evaluate to what extent the identified similarities align with perceived similarities by the musicological community.

6.5 Results

This section describes the results of quantitative and qualitative evaluation.

Quantitative Evaluation

Table 6.2 shows the baseline results with raga datasets combinations and aforementioned general English corpus. Table 6.3 compares the accuracy of different *min-count* with different dataset combinations taking the base word vectors as Google and general English Corpus (*Gen*

Datasets	Score	
	BL-1	**BL-2**
sganga + wiki	0.0813	0.0945
sganga + wiki + meer	0.1063	0.1023
sganga + wiki + meer + parrikar	0.1016	0.1117

Table 6.2: Results: Baseline experiments (score: obtained with the described evaluation method, *min-count*=1).

English) with our approach. The experiment with all the datasets combined shows best accuracy and *min-count*= 1 gives better accuracy through out.

Experiment	Base word vectors	Score		
		min-count=1	**min-count=2**	**min-count=3**
sganga+wiki	Google	0.0867	0.0695	0.0664
sganga+wiki+meer		0.1141	0.0828	0.0844
sganga+wiki+meer+parrikar		0.1172	0.1039	0.0977
sganga+wiki	Gen English	0.1016	0.0648	0.0672
sganga+wiki+meer		0.1078	0.0945	0.0977
sganga+wiki+meer+parrikar		0.1140	0.0992	0.1047

Table 6.3: Results of our approach with different base word vectors and *min-count* (score: obtained with the described evaluation method).

Compared to the baseline experiments (Table 6.2), our approach shows better accuracy for all the dataset combinations. BL-1 performance is affected due to insufficient text. Baseline BL-2 is better comparable with the experiments taking *Gen English* as base word vectors in Table 6.3, since it uses this same general English data combined with raga datasets. Our approach has clear performance improvement over BL-2 with *Gen English* experiments. Baselines are executed with *min-count*=1. One key observation here is, word vectors generation done with *min-count*=1 adversely affects the common word vectors which in turn affects the raga word vectors (as in BL-2). Our method rectifies this with the two-step word vectors training with

variable *min-count* (*min-count*=10 for generating base word vectors and *min-count* < 4 in the incremental generation step).

6.5.1 Qualitative Evaluation

Qualitative evaluation is done with the help of 2 Hindustani musicians. One is trained professional musician in Hindustani music and the other is an amateur with basic understanding of Hindustani music. The main analysis produced here is done by the professional musician and the amateur musician did overall analysis confirming many of the former's observations. The musicians were provided with a t-SNE visualization (Van der Maaten and Hinton, 2008) of the raga word vectors and text listing 10 closest (based on cosine similarity) ragas to each raga. This visualization with some of the observed similarities magnified, is shown in Figure 6.2. The musicians evaluated for all 3 dataset combinations.

sganga+wiki+meer+parrikar and *sganga+wiki +meer* are found to be superior to *sganga+wiki* w.r.t the similarities captured. Among them, *sganga+wiki+meer+parrikar* has more relations captured. This observation has a direct correlation with the results in Table 6.3 proving the validity of the quantitative evaluation method. Overall many of the close ragas in the visualization belong to the same *thaat* in all the 3 cases. Since *thaat* based similarity corresponds to melodic similarity and is considered to be prime among the other raga attributes, the identified similarities sound meaningful.

Ragas from Kafi, Asavari, Bhairavi *thaats* are closely placed. These *thaats* have a minor scale with a change in a single note. A few ragas which are perceived as similar and have less commonalities w.r.t raga attributes are identified similar (eg. Patdeep, Nanad and Tilak Kamod, Komal Rishabh Asavari). These are perceived similar because of the similarities in their characteristic phrases. Between Patdeep and Nanad, the main characteristic phrase differ by a note. Certain pentatonic ragas *viz.*, Bhoop, Gunkri, Arabhi, Shivaranjani and Malkauns are found close to each other (refer (b) in Figure 6.2). Certain clusters are observed with ragas similar w.r.t group (eg. Malhar group ragas are found together) and *prahar*. (c) in Figure 6.2 shows 2 different clusters of malhar group ragas. (d) shows that the *misra ragas* Lalit bhatiyar, Jogkauns, Nat Bihag, Nat Bhairav, Asa Mand, Puriya Kalyan, Puriya Dhanashree, Bhupal Todi, Basanti Kedar are placed closer. Some of the ragas adopted from Carnatic music are identified closer to each other.

From the 2D visualization, the amateur musician made an interesting observation that

many of the ragas belonging to the same *thaat* lie in a diagonal strip on the plane, with a negative slope. Also frequently sung ragas are found closer to the center of the visualization, where as rare ragas are seen away from the center.

Along with these positive observations there are a few undesired similarities appearing. A few pair of ragas which are supposed to be far apart are seen closer (eg. (Basant, Bageshri), (Poorvi, Hameer)). Most of the clusters of similar ragas identified, are found to be grouped together based on one of the attributes, which the musicological community also considers relevant for that grouping. The low number of invalid similarities compared to missing valid similarities, shows high precision of the system.

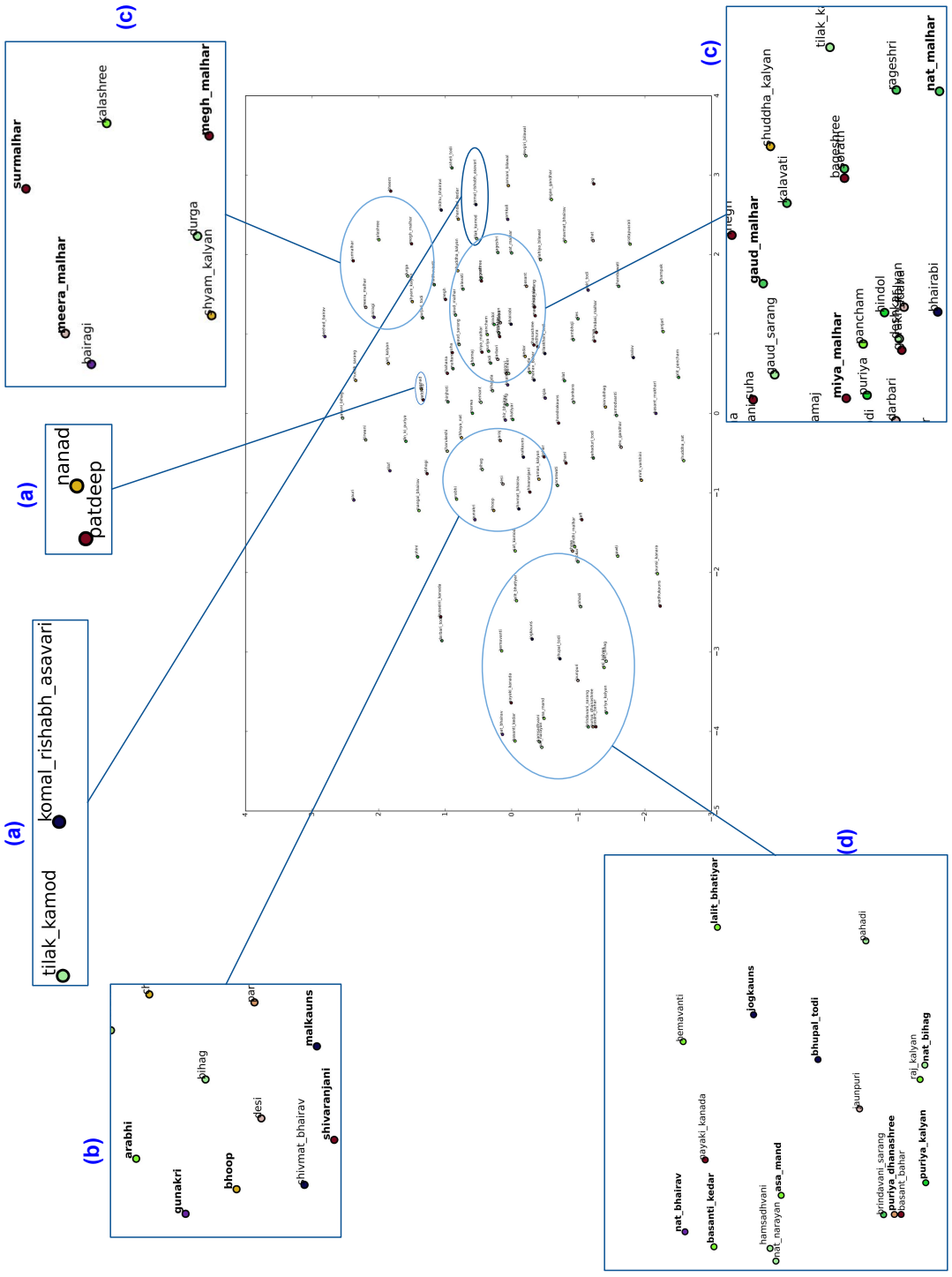


Figure 6.2: TSNE visualization of raga similarity with valid similarity clusters shown in magnified windows. Relevant raga names in a magnified window are shown in bold.

6.6 Applicability of the Approach to Other Domain-Specific Datasets

So far, we performed our experiments on Hindustani raga related text. To evaluate our approach better, we also experimented with generating word vectors for 2 different datasets *viz.* US political tweets (Joshi et al., 2016) and CoNLL 2003 NER (Tjong Kim Sang and De Meulder, 2003) datasets. With this evaluation, we try to see the applicability of our approach in learning word vectors for domain-specific words in any domain with insufficient data. In these experiments, for a set of words defined as dataset-specific words in the datasets, we learn the word vectors with our approach. Clustering based evaluation is employed to see how good these dataset-specific words get clustered according to their known classes.

From US political tweets, position words belonging to 6 different political issues (*eg. gun laws, immigration, insurance etc.*) are considered as dataset-specific words. From CoNLL 2003, all the entities belonging to 4 named entity classes are taken as dataset-specific words. Here the evaluation is done by analyzing how well these words get clustered according to their known classes. In the US political tweets dataset, the position words are expected to be clustered according to the issues they belong to and in CoNLL 2003 dataset the entities are expected to be clustered according to their named entity classes.

Dataset	**Experiment**	**Homogeneity**	**Completeness**	**V-measure**	**Rand**	**Mutual Information**
US political tweets	BL-1	0.4185	**0.5439**	0.4730	0.1619	0.2802
	BL-2	0.3124	0.3519	0.3310	0.1425	0.1506
	Our Approach	**0.4583**	0.5069	**0.4813**	**0.1719**	**0.3215**
CoNLL 2003 NER dataset	BL-1	0.0088	0.0088	0.0088	0.0201	0.0085
	BL-2	0.0025	0.0197	0.0045	0.0088	0.0022
	Our Approach	**0.0120**	**0.0250**	**0.0163**	**0.0464**	**0.0117**

Table 6.4: Results of clustering evaluation with 2 different datasets (best results shown in bold).

Word vectors for these words are learned using the baseline approaches and our approach. K-means clustering is employed, defining k as the number of word categories. Our approach is compared with the baselines by evaluating the clustering of words using a set of standard metrics

for clustering. Table 6.4 shows that our approach performs better for both the datasets w.r.t all the metrics, with a single exception. Along with affirming the applicability of our approach for any dataset with rare words, this evaluation also helped to perform a strong validation with tasks having solid ground truth.

6.7 Summary

The approach discussed in this chapter for extracting raga similarities from text discussions learn word vectors for the raga words. To learn word vectors efficiently from the available scarce dataset, we proposed a modification to Mikolov's approach for learning word vectors. Quantitative and qualitative evaluations substantiates that the similarities between the raga word vectors emulate perceived raga similarities. Comparison with the baseline methods confirms the utility of this method to learn representation for raga words. Further validation of the method is done by clustering evaluation performed with two other datasets. This confirms the practicality of the approach to learn word vectors for any dataset-specific words. As future work, we plan to improve this method by considering the relative importance of context words variably, as certain word categories are more important for this task.

Chapter 7

Information Extraction from Music Discussion Forums: Relevance of Coreference Resolution

In the context of MIR, information extraction from text helps in extracting meta information from the large available online content to enrich the music information knowledge base along with information extracted from audio content. Some discussion forums and blogs are rich sources of information pertaining to Indian classical music. While serving as the common platform for discussions, many of the discussion forums turn into a one-stop source for versatile information in their respective domains. This chapter discusses the task of coreference resolution, motivated by the requirement of resolving anaphoric mentions for improved knowledge extraction from discussion forums on Indian classical music. The task of coreference resolution resolves the anaphoric mentions against the entities it refer to. A mention is a word or a phrase in the text referring to an entity (Doddington et al., 2004). Even though coreference resolution is a widely researched problem, we investigate on this problem considering the nature of the text in discussion forum and the domain of music. Among the classical music followers in India, the Carnatic music community is more involved in web based discussions and information dissemination. Rasikas.org (ras, 2017) is one among the prominent discussion forums pertaining to Carnatic music topics comprising of ragas, *talas*, artistes, etc. Instead of a discussion forum on Hindustani music for our study, we consider Rasikas.org which is a discussion forum on Carnatic music because of its popularity and the amount of content available. Also, both the traditions have similar concepts and terminologies.

In this chapter, we discuss our approach for coreference resolution on posts from Rasikas.org. The commonly used mention-pair paradigm (Aone and Bennett, 1995) is followed for modeling the problem. Here the focus is on feature engineering and mention-pair classification. Our dataset has taken posts from forums of different sub-topics on Carnatic music from Rasikas.org. As in any data-driven approach, the selection of features is imperative here as well. We design new features and apply modifications to certain features to cater to the requirements of the dataset. Further, we discuss our experiments on mention pair classification with different classifiers. Lack of annotated data in the initial phase of this research motivated the use of Bayesian network as classifier with hand-engineered network structure. With the extension of our dataset with more annotated data, Bayesian network was no longer the best. Considering the diverse knowledge required for the coreference classification, we also discuss how a hybrid approach can help mention pair classification.

The focus of this chapter is on introducing the problem modeling and feature engineering along with the classification approaches. A detailed description of the dataset is given in Section 7.1. Our approach to modeling the problem is discussed in Section 7.2. The discussion on features including the novel features introduced and modifications tried are in Section 7.3. Section 7.4 discusses our approach to mention detection. The classification approaches with Bayesian network and hybrid approach are discussed in Sections 7.5 and 7.6. This is followed by experiments and results in Section 7.7

7.1 Dataset: Rasikas.org

The coreference annotated dataset contains forum posts from Rasikas.org. This is a prominent discussion forum on Carnatic music. The main topics of discussion in the forum includes raga (Bhagyalekshmy, 1990), *tala* (rhythm), *vidwans & vidushis* (musicians), *vaggeyakaras* (composers), *kutcheri* (concert), reviews & recordings, album reviews, etc. This forum is a rich source of information, including music listeners' opinions in the mentioned topics. Table 7.1 shows the details of this dataset.

Each thread in a forum topic discusses on a sub-topic in that category. For instance, in the forum on musician, each thread deals with a particular musician. Posts in forums are written in informal language with pronominal and alias mentions referring to the main topic of discussion or to another related entity mentioned in the discourse. Efficient extraction of relation

Forum	#Posts	#Sent.	#Mentions
Raga & Alapana	300	2093	3631
Vidwans & Vidushis	587	3045	10884
Vaggeyakaras	325	2339	4421

Table 7.1: Details of annotated posts.

is dependent on finding the exact antecedent of pronominal and nominal mention, when it refers to another entity. It is commonly observed that the main topic of a post is referred frequently. A post from the forum is given below. Coreferent mentions are marked with the same color.

Sri Ragam is the asampoorna mela equivalent of K Priya acc to MD's school. Thyagaraja gave life to K.Priya with his excellent compos, where as MD never touched this raga. In Sri ragam we have plenty of compos by the trinity incl the famous Endaro Sri Ranjani is a lovely janya of K Priya with plenty of compos by both T & MD.

Table 7.2 shows the mention type statistics of annotated mentions in Rasikas dataset. The annotation also includes singleton mentions leading to relatively large number of proper nouns within annotated mentions. Singleton mentions are mentions which are not corefered by any other mention in the document, contributing to chains with single mention in it.

Mention Type	Count
Proper nouns	11819
Pronouns	5664
Definite phrases	863
Indefinite phrases	236
Demonstrative phrases	354
Total	**18936**

Table 7.2: Mention type statistics of Rasikas dataset

The text content is very much similar to asynchronous or synchronous online communication style. The similarity of the forum content with spoken language makes it distinct. The

content comprises mixture of written and spoken discourse, reflecting the orality of online communication styles. This is attributed also with a few grammatical errors, less structuring and spelling discrepancies especially with the named entities. Each forum post is a short discourse text comprising 4-5 sentences on an average. The short discourse of text in these forums with distinct characteristics makes it imperative to have a dedicated coreference resolution system designed for this dataset.

The manual annotations are done by an annotator with proficiency in English. GATE (Cunningham et al., 2011), a package with annotation capabilities is employed to help the annotator. GATE also has solutions for different stages of text processing. Prior to annotation, sentence tokenization and word tokenization are done automatically with the functionalities available with GATE. The annotation tool having color coding features make the visual verification easier while annotating. The annotations are further processed to convert it to CoNLL format. This is a tabular format with one line per token. CoNLL format is a widely accepted format for coreference annotated dataset because of its capabilities in incorporating different layers of information including named-entity class, constituency parse tree, dependency parse tree etc (Pradhan et al., 2012). The annotations done by an annotator is verified by the author of this thesis. Along with the mentions which are part of a coreferent chain, some of the relevant singleton mentions are also annotated for a document.

7.2 Our Approach

We follow a supervised learning approach, formulating the problem using mention-pair model (explained in Chapter 2, Section 2.3.4), where the coreference classification decision is made for the mention pairs followed by clustering, forming distinct chains of coreferent mentions (Recasens and Hovy, 2009; Aone and Bennett, 1995; McCarthy and Lehnert, 1995). In the classification step, features for mention pairs are computed. After training is done with the mention pair instances from the training documents, mention pair instances from test documents are classified as coreferent or not. Coreference clustering forms coreferent chains of mentions using the mention pairs classified as coreferent. For clustering, we use best-first clustering (Ng and Cardie, 2002b). For an anaphoric mention, best-first clustering takes the candidate classified with highest confidence as the antecedent, from the candidate antecedents. Mentions which are classified as coreferent with the anaphoric mention in the mention pair classification step are

taken as candidate antecedents. The probability estimate value associated with the classification is taken as the classification confidence score.

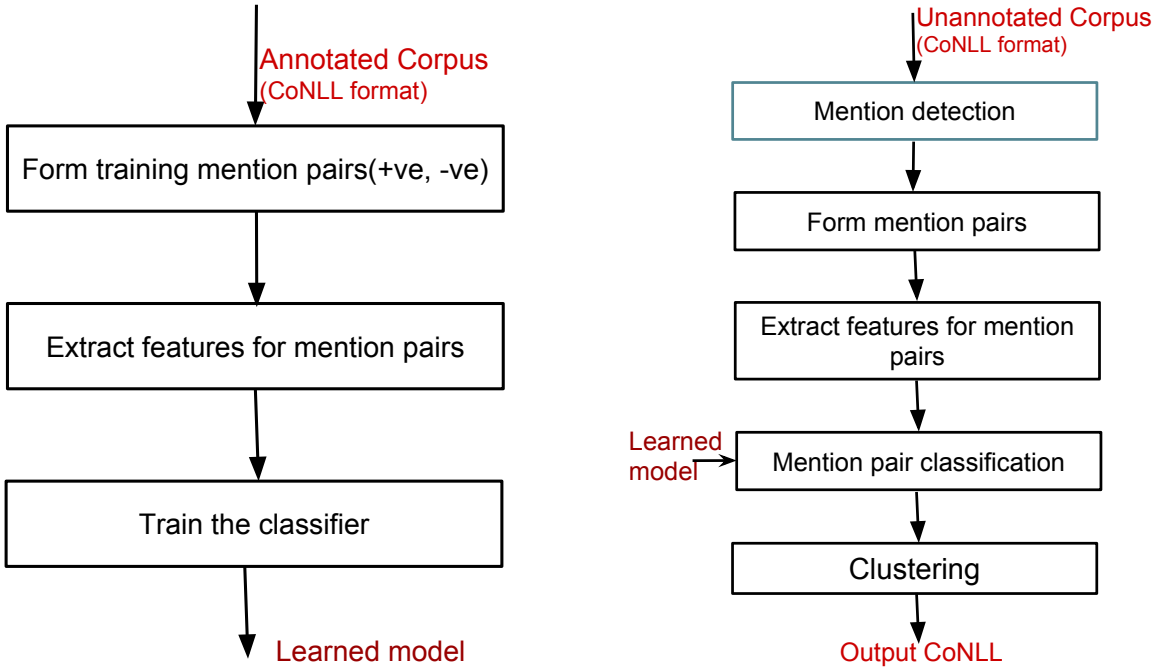


Figure 7.1: (a) Training (b) Testing steps in our approach following mention-pair model

Figure 7.1 gives an overview of the steps in training and testing of our approach. The training step forms coreferent mention pairs (positive instances) with mentions within coreferent chains and non-coreferent mention pairs (negative instances) with mentions from different coreferent chains for training the mention pair classifier. Training depends on gold mentions (annotated mention boundaries), whereas testing depends on automatically detected mentions identified by automatic mention detection step. For some of our experiments, we have reported results with gold mentions as well. Testing mention pairs are formed with all possible combinations of mentions available in a test document. The number of such mention pairs formed are controlled with certain heuristics. As mentioned, testing involves mention pair classification followed by clustering.

7.3 Knowledge Source for Coreference Resolution

Features are computed for a mention pair. Each mention pair comprises of a candidate antecedent mention and an anaphoric mention. Design of the features are influenced by the do-

main considerations. Some novel features are introduced and some of the existing features are modified accordingly. We make use of a subset of conventional features including the features described in Soon et al. (2001). Here we explain the selected conventional features and the data-specific modification if any. All features are explained for a mention pair (m_i, m_j), where m_j is the anaphoric mention and m_i is the candidate antecedent.

String match (STR_MATCH): This feature checks for string similarity between the mention strings. Different from the existing implementations of this feature, we use fuzzy string matching instead of strict string matching. We use an edit distance method, Levenshtein distance (Levenshtein, 1966) for fuzzy string matching. This is to handle proper nouns spelled variably at different places within a forum post. For example, the raga name 'Huseni' is also spelled as 'Huseini'. This feature is true, if the Levenshtein distance between the mentions is within a defined threshold. It is also observed that English as well as Indian honorifics such as 'Mr', 'Mrs', 'Shri', 'Smt', are used along with person names. To make the string similarity checking independent of this, pre-processing is done to remove these addressing terms.

Alias (ALIAS): This feature checks if one mention is an alias for the other. If one mention string is the abbreviation of the other or a part of the other mention string, it is considered to be an alias of the other. The aforementioned Levenshtein distance is applied here to check for the partial match, comparing the shortest among the mentions with each token of the longer mention. Here also we do the pre-processing to eliminate the honorifics from the mention string.

Definite noun phrase (m1_DEF, m2_DEF): This feature checks if the mention is a definite noun phrase. A noun phrase starting with *the* is a definite noun phrase (eg. the song). m1_DEF does this check for the antecedent mention (m_i) and m2_DEF for the anaphoric mention (m_j).

Indefinite noun phrase (m1_INDEF, m2_INDEF): This feature checks if the mention is an indefinite noun phrase. A noun phrase starting with *a, an* is an indefinite noun phrase (eg. a performance). This feature is also computed for both the mentions independently.

Demonstrative noun phrase (m1_DEM, m2_DEM): This feature checks if the mention is a demonstrative pronoun. A noun phrase starting with *this*, *that*, *these* and *those* is a demonstrative noun phrase (eg. this raga). This feature is also available for both the mentions independently.

Pronoun (PRN): This feature checks if the anaphoric mention is a pronoun.

Proper noun (m1_PRPN, m2_PRPN): This feature checks if the mention is a proper noun. A mention is considered for this check, only if it is not a pronoun, definite phrase or a demonstra-

tive phrase. If a mention string starts with a honorific term as mentioned above, it is considered as a proper noun. If all the letters of the mention are in caps or the mention string is title cased (first letter of all the words in capital), then also the mention is taken as a proper noun. This feature is available for both the mentions independently.

Same sentence (SAME_SENT): This feature checks if both the mentions are in the same sentence.

First sentence (FIRST_SENT): This feature checks if the antecedent mention (m_i) is in the first sentence of the forum post. In a forum post, most of the discussion pertains to the main topic which is likely to appear in the first sentence of the post.

Sentence distance (DIST_SENT): This features computes the distance between mentions in a pair, in terms of number of sentences present between them.

Word distance (DIST_WORD): This features computes the distance between mentions in a pair, in terms of number of words present between them.

Addressing mention (ADDR): This feature checks if the first mention in the pair is addressing someone. The mention *John* in *John, Where are you?* is an addressing mention. This feature is relevant for a discussion forum because of the presence of sentences with addressing other forum users.

First person personal pronoun (FIRST_PER_PRN): This feature checks if the mention is a first person personal pronoun. First person personal pronoun includes *I, me, mine*. This feature is relevant considering the frequency of usage of first person personal pronouns in the forums.

Second person personal pronoun (SECOND_PER_PRN): This feature checks if the mention is a second person personal pronoun. Second person personal pronoun includes *you, your*. This pronoun type is also found often in forum posts.

Considering the nature of the dataset, number agreement and gender agreement features are difficult to be extracted. The number agreement feature is not relevant for this dataset, since there are not many plural entities. Even though gender feature is important, the existing resources are not helpful for identifying the gender of Indian names.

7.3.1 Grammatical Role Features

Though the discussed features are significant for showing the coreferent characteristics of a mention pair, features involving the grammatical role of a mention in a discourse are also important to coreference identification. In a short discourse where the mentions lie in close vicin-

ity, the grammatical role is an important player in deciding coreference, when compared to long discourse having coreferent mentions far apart. Table 7.3 describes the basic grammatical role features.

Feature	**Description**
First mention subject (m1_SUBJ)	True, when m_i is a subject of any verb in the sentence
Second mention subject (m2_SUBJ)	True, when m_j is a subject of any verb in the sentence
First mention object (m1_OBJ)	True, when m_i is an object of any verb in the sentence
Second mention object (m2_OBJ)	True, when m_j is an object of any verb in the sentence

Table 7.3: Basic grammatical role features

Apart from analyzing whether a mention in the pair is a subject or an object in a sentence as discussed in literatures (Ng and Cardie, 2002b; Lappin and Leass, 1994; Recasens and Hovy, 2009), we also analyze the role of other mentions coming in between the mentions of the pair under consideration. This helps to figure out the existence of any other potential antecedent for the anaphora in the mention pair (m_i,m_j). The existence of a potential antecedent should decrease the probability of the mention pair considered, to be coreferent. The role of a mention is determined with the help of dependency parse of a sentence. The coreferent relation between two mentions is also dependent on other mentions occurring around the mentions under consideration. So we designed a few other features to capture the behavior of other neighboring mentions, in order to supplement or weaken the coreferent relation between the mentions in the pair under consideration.

Subject mention between (SUBJ_BET): This feature is true when there is another mention in between m_i and m_j, having subject dependency relation to a verb in the occurring sentence. This feature is intended to reduce the probability of a mention pair becoming coreferent when there is a potential candidate present in between. Consider this example

`KR Kedaranathan` is the topic of this thread. Smt Meera Kedaranathan, wife of late Sri KR Kedaranathan passed away on the evening of 12 January 2014. `She` was also a disciple of Semmangudi Srinivasa Iyer and has trained a number of students in music.

Here the pair (`KR Kedaranathan, She`) is not coreferent. The existence of `Smt Meera Kedaranathan` make it clearly non-coreferent. This feature helps to reduce the chances of `KR Kedaranathan` getting classified as coreferent with `She`, because of the presence of the subject `Smt Meera Kedaranathan` in between.

Subject mention associated with root verb between (ROOT_SUBJ_BET): The root verb is the main verb in a sentence or a clause. This feature is a complement to the previous one, checking for existence of a mention between m_i and m_j having a subject dependency relation with the root verb of the sentence. Such a mention has higher probability of being antecedent to the current anaphoric mention.

First mention subject of root verb (MEN1_ROOT_SUBJ): This feature checks for whether the first mention in the pair is associated with the root verb in the occurring sentence. This increases the chance of this mention being referred in the subsequent sentences more often.

7.4 Mention Detection

Mention detection is an important step to identify the potential mentions in a text prior to performing coreference resolution (Soraluze et al., 2012). In a realistic setting, while the training step depend on gold mentions from the annotated training data, the testing takes the predicted mentions identified by the mention detection step. The accuracy of mention detection is very critical to the accuracy of the coreference resolution system (Stoyanov et al., 2009; Hacioglu et al., 2005; Zhekova and Kübler, 2010). Many of the existing mention detection approaches depend on constituency parsers to extract the noun phrases from a sentence. Further, appropriate filters are applied to refine the noun phrases according the general nature of mentions in the dataset (Kummerfeld et al., 2011; Haghighi and Klein, 2010; Broscheit et al., 2010; Chen and Ng, 2012).

For our dataset, an approach based on constituency parser is likely to perform bad, due to the presence of grammatical errors and ill-structured sentences. Also, the annotated mentions in our dataset do not include long noun phrases. This exclusion is driven by the nature of potential relations extractable for the task of relation extraction from the forum. Taking these into account, all the noun phrases identified by a constituency parser are not potential mentions in our dataset. A rule-based chunker is deployed to extract mentions, limiting the extraction

to predefined part-of-speech tag patterns which are identified from observations on annotated mentions. To improve the accuracy of mention detection, certain heuristics are also applied to correct the omissions. Errors in POS tagging cause some pronouns, nominals and proper nouns (especially Indian terms) to get unidentified. Post processing following the pattern based identification, checks for missed pronouns and proper nouns. The following description explains the problems of the mention detection approach and the corresponding heursistics implemented to resolve it.

- Certain pronouns are found unidentified. The heuristic to solve it checks if some pronouns are missed. This check is done with the help of a gazetteer of pronouns.

- Some proper nouns are found unidentified. The heuristic to solve it depends on an English dictionary. If a word does not appear in the English dictionary, it is most likely to be a proper noun.

- There are cases where verbs are identified as proper nouns. This is corrected with the help of a secondary POS tagger. A secondary POS tagger; spacy[1] is employed. Spacy is a deep learning based NLP library which is identified to be good in disambiguating certain tags. All the proper nouns are subjected to a double check with this secondary POS tagger.

- The words having possessive endings are clearly nouns (eg. songs) which are potential mentions. There is a check to identify if words with possessive endings are not identified as mentions because of the problems with the POS tagging.

7.5 Bayesian Network for Small Dataset

This investigation was done in the initial phase of this research on coreference resolution, when the available annotated dataset was small. To deal with data insufficiency we try with Bayesian network for mention pair classification. This is evaluated with a defined network structure designed to capture known dependencies between the features. In this approach, we employ a simple network structure designed with the help of prior knowledge on the dependencies between the features. Antal et al. (2004) claims that, when the data is scarce and we know the dependencies between features, Bayesian network performs better. We experiment with Bayesian

[1] http://spacy.io

network structure with hand-engineered dependencies between the features. We observe that, the performance of Bayesian network with hand-engineered network structure performs better compared to other classifiers. On availability of sufficient annotated data, we observed that Bayesian network no longer performs better compared to other classifiers. We propose this as a classification approach for building a coreference resolution system when the annotated dataset is small.

7.5.1 Bayesian Network

A Bayesian network is a probabilistic graphical model representing dependencies between a set of random variables. The nodes in a Bayesian network represent the random variables and the directed arcs between them represent the dependencies between the variables. Given a Bayesian network with n nodes ($X_1, X_2, ...X_n$), the joint probability distribution with specific values assigned to each random variable is represented as $P(X_1 = x_1, X_2 = x_2, ...X_n = x_n)$. Using chain rule this is factorized as

$$\begin{aligned} P(x_1, x_2, ...x_n) &= P(x_1) \times P(x_2|x_1) \times .... \times P(x_n|x_1, x_2...x_{n-1}) \\ &= \prod_i P(x_i|x_1, x_2, ...x_{i-1}) \end{aligned} \tag{7.1}$$

Based on Markov property of Bayesian networks, each variable is conditionally independent of its non-descendants, given its parents. The probability of a particular node is conditional only on the probability of its parent nodes (Korb and Nicholson, 2010). So this can be factorized as

$$P(x_1, x_2, ...x_n) = \prod_i P(x_i|Parents(x_i)) \tag{7.2}$$

When Bayesian network is employed for classification task, we have a random variable for class (C) along with the other random variables representing the features ($A_1, A_2, ...A_n$). Here the joint distribution can be represented as $P(C, A_1, A_2, ...A_n)$. For a data instance with features $(a_1, a_2.....a_n)$, the Bayesian network classifier based on a specific network structure returns the class c that maximizes the posterior probability $P(c|a_1, a_2.....a_n)$ (Friedman et al., 1997).

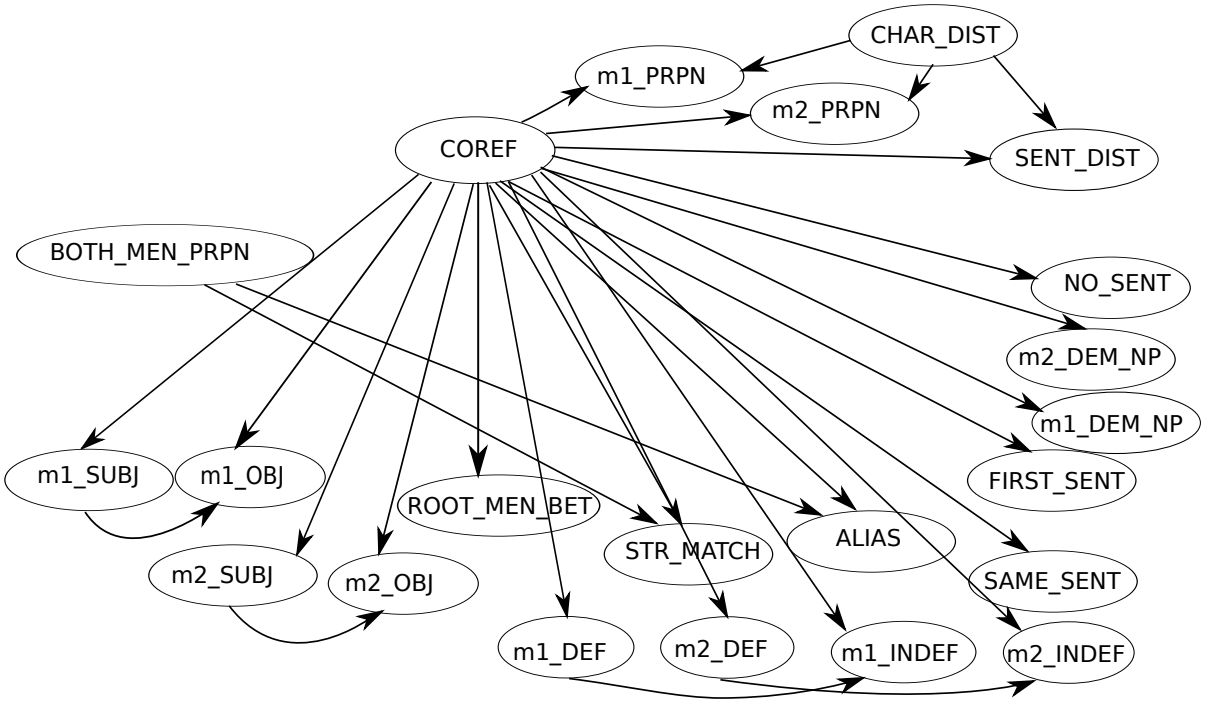


Figure 7.2: Hand-engineered Bayesian network

7.5.2 Bayesian Network Design

Figure 7.2 shows the dependencies between the features. The node '`Coref`' represents the random variable for the coreference class. Here we experiment with most of the features discussed in Section 7.3. Following is a description on some of the important relations encoded into the network. STR_MATCH and ALIAS features are influenced by BOTH_MEN_PRPN, because these features are relevant only when both the mentions are proper nouns. A mention can take the grammatical role of either subject or object. This results in a relation between M*_SUBJ and M*_OBJ. Same is the reason for the relations between M*_DEF and M*_INDEF. When one of the mentions is a proper noun, the decision of coreference also depends on the distance between the mentions. This is realized by a relation from CHAR_DIST to M*_PRPN.

7.6 Hybrid Approach

The human process of coreference resolution involves heterogeneous classification tasks depending on the type of mention to be resolved. The knowledge applied for resolving a proper noun anaphoric mention is different from the knowledge applied for resolving a pronominal anaphoric mention. These observations led to a hybrid approach combining rule-based ap-

proach and machine learning (ML) based approach, segregating the classification of mention pairs which are better classified by rules.

Our hybrid approach combines a rule based sieve with a machine learning based classifier. The hybrid approach is motivated by the domain specificities and the amount of annotated data. The rule based sieve primarily takes care of mention pair classification requiring lexical similarity based features. The requisite features for such mention pairs is limited and can be better classified with a set of rules. Segregation of these classifications help the machine learning based classifier (the latter part of the hybrid approach) to learn better for other mention pair categories.

While training the ML classifier of the hybrid system, the mention pairs in which both the mentions are proper noun are excluded from the training data. These data instances are very likely to be classified by the rule-based sieve. As mentioned, the main reason for the mentions in these mention pair instances to be coreferent is the lexical similarity between the mention strings. Even though other features are irrelevant, making this a part of the training data to ML classifier may affect the classification of data instances which depend on other knowledge sources. So this exclusion helps to improve the ML classifier in classifying the data instances passed to the ML classifier in the hybrid system during testing.

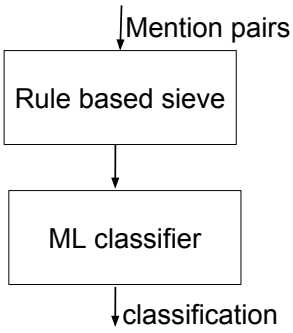


Figure 7.3: Hybrid approach

There are a few existing hybrid approaches for coreference resolution. Chen and Ng (2012) combines the strengths of rule based and learning based methods for English, Chinese and Arabic. Nilsson (2010) proposed a hybrid approach for Swedish text, understanding the differentiation between the subtasks involved in coreference resolution. The hybrid approach by Hendrickx et al. (2007) combines rule based filtering with machine learning based method. Linguistically motivated positive and negative filters form the rule based filter. In the hybrid solution proposed for Hindi, rule based module make use of dependency structures to resolve

coreferences and a decision tree based classifier classifies the unattended mentions using grammatical and semantic features (Dakwale et al., 2013).

7.6.1 Coreference Evaluation

Coreference evaluation checks for matching between the coreferent chains formed by the experimental system and the coreferent chains in the gold standard corpus. At present, there are three widely accepted coreference metrics, viz. MUC, B^3 and CEAF . All these metrics report results in terms of precision, recall and F-measure. CoNLL shared task on coreference resolution in 2011 introduced CoNLL score to combine these F scores by taking the average of F-scores of these metrics. MUC (Vilain et al., 1995) compares the system response with the gold standard input by counting the minimum number of links to be inserted or deleted between the mentions. MUC does not evaluate the system's ability to separate out the singleton mentions (mentions not belonging to any coreferent chain). B^3 (Bagga and Baldwin, 1998) overcome this shortcoming by computing the accuracy based on mentions instead of links. But B^3 fails in handling the mentions which are there in system response but not in the gold standard. The metric CEAF (Luo, 2005) is proposed to solve the the problem of B^3 counting the entities more than one time. CEAF applies a similarity metric to compute the best alignment between a set of mentions (Cai and Strube, 2010).

7.7 Experiments & Results

In this section, we discuss the results of coreference resolution with the classification approaches discussed. Before this, we also discuss the results of our mention detection approach and present an evaluation of the features for coreference.

7.7.1 Mention Detection

Mention detection depends on defined POS tag patterns. We use the regular expression based chunking functionality available with NLTK[2] for this. POS tagging is done by Stanford POS tagger (Toutanova and Manning, 2000). Spacy toolkit[3] is used as the secondary POS tagger to resolve the confusing tags.

[2]Natural language Toolkit: http://www.nltk.org/

[3]http://spacy.io

ID	Pattern	Example
1	<DT>?<NN>*<NNS>*<NNP>+<NN> *<NNS>*<FW>*<NNPS>*<POS>*	*a Carnatic composition, Muthuswami Dikshithar, the asapoorna mela equivalent*
2	<NN>*<POS>	*raga's, performer's*
3	<DT>*<NNS>*<NN>*<FW>+<NN> *<NNS>*<NNPS>*<POS>*	*the raga performance, keertana*
4	<PRP>|<PRP$>	*he, she, it, his*

Table 7.4: Mention detection: POS tags patterns

Table 7.4 shows some of the defined patterns with corresponding examples. Pattern 1 identifies all mentions involving a proper noun. Pattern 3 is also similar to this, but to identify mentions having a word tagged as foreign word (FW). Pattern 2 identifies possessive nouns and pattern 4 identifies all pronouns. Adding more patterns to improve the mention detection recall may lead to increase in false positives. Therefore instead of adding more patterns, certain heuristics are also employed to detect more mentions. Table 7.5 shows the accuracy of mention detection.

Precision	Recall	F1
65.89	78.11	71.48

Table 7.5: Mention detection accuracy (in %)

7.7.2 Bayesian Network for Small Dataset

In this experiment we see the effectiveness of Bayesian network with hand-engineered network structure over other classifiers for a small dataset. This is a subset of the larger dataset mentioned in Table 7.1. Details of this smaller dataset is shown in Table 7.6.

We compare Bayesian network results with Naive Bayes, SVM (RBF) and a simple feed-forward neural network. We also compare with Bayesian network with learned network structure. We use libpgm[4] python package for implementing Bayesian network. We use scikit-learn

[4]http://pythonhosted.org/libpgm/

Forum	#Posts	#Sent.	#Mentions
Raga & Alapana	143	893	2091
Vidwans & Vidushis	180	1219	2749

Table 7.6: Details of small dataset. (#Posts= No. of posts #Sent= No. of sentences in the forum. #Mentions= No. of annotated mentions)

(Pedregosa et al., 2011) for Naive Bayes and libSVM[5] for SVM implementations. The selection of parameters for SVM classifier is done with a small development set taken from the training data. We use Keras for the neural network implementation.

Experiments	**MUC**			**B^3**			**CEAFe**			**CoNLL Score**
	P	R	F	P	R	F	P	R	F	
Naive Bayes	31.79	30.48	30.81	42.88	47.84	44.95	28.33	51.46	36.19	37.32
SVM (RBF)	42.25	54.59	47.36	42.07	58.64	48.97	37.36	58.03	45.06	47.13
Neural net	37.13	63.21	46.72	38.32	64.36	48.03	42.88	57.26	48.90	47.89
Bayes net (Learned)	32.11	28.69	27.86	43.85	47.49	44.35	28.43	52.37	35.88	36.03
Bayes net (Hand-engineered)	45.48	57.88	50.77	42.44	60.48	49.87	36.77	57.77	44.64	**48.42**

Table 7.7: Results with small dataset. P:precision, R:recall, F:F-measure. Learned: Learned network structure Hand-engineered: Hand engineered network structure

Table 7.7 reports results in MUC, B^3 and CEAFe metrics. CoNLL score averages F-scores of these metrics. Bayesian network with hand-engineered network structure outperforms other classifiers on this small dataset. The accuracy of neural network is close to this. Even though neural network gives the best recall with all the metrics, the low precision leads to lower F-measure. Bayesian network with learned network structure shows the worst performance.

For all other experiments including feature evaluation, we use the extended dataset mentioned in Table 7.1.

[5]http://www.csie.ntu.edu.tw/ cjlin/libsvm

Experiments	**MUC**			**B^3**			**CEAFe**			**CoNLL Score**
	P	R	F	P	R	F	P	R	F	
Naive Bayes	37.53	37.58	37.55	50.17	53.08	51.58	39.52	50.65	44.39	44.50
SVM (RBF)	50.25	59.48	54.47	51.64	63.60	57.00	49.86	57.78	53.53	55.00
Neural net	54.50	61.69	57.87	54.30	65.27	59.28	50.45	60.38	54.97	57.37
Bayes net (Hand-engineered)	51.67	56.28	53.88	50.52	61.67	55.54	45.78	56.88	50.73	53.38

Table 7.8: Results with extended bigger dataset. P:precision, R:recall, F:F-measure. Learned: Learned network structure Hand-engineered: Hand engineered network structure

7.7.3 Feature Evaluation

We analyze the feature importance of all the features using decision tree. Gini importance obtained from training the decision tree is utilized for identifying feature importance. Gini importance is derived from Gini index. Gini importance captures the relevance of a feature by computing the separation obtained by selecting that feature (Venkataraman et al., 2010). Figure 7.4 shows the feature importance in the descending order (from top).

As reported in other existing literature, STR_MATCH is the most discriminating feature. ALIAS which is another important lexical feature is also identified as one of the important features. Grammatical role features are also found relatively of higher importance. OBJ_m1 is next to STR_MATCH, and SUBJ_m2 and SUBJ_MEN_BET are within the first half. But ROOT_SUBJ_BET and MEN1_ROOT_SUBJ are identified as features of no importance. These grammatical role features with Gini importance 0 are not considered for the subsequent experiments.

Table 7.9 shows the results in CoNLL score with different categories of features. Here we experiment with 3 broad categories of features with the 2 best-performing classifiers. The first category (CAT-I) contains only the lexical features (STR_MATCH, ALIAS, SAME_HEAD), second category (CAT-II) includes most of the other features pertaining to other categories except for the grammatical role features. To check the relevance of grammatical role features for coreference resolution in short texts, we experiment them as a different category (CAT-III). CAT-III includes SUBJ_m1, SUBJ_m2, OBJ_m1, OBJ_m2, ANY_MEN_BET, SUBJ_MEN_BET. Each category features are incrementally added in these experiments with the best performing

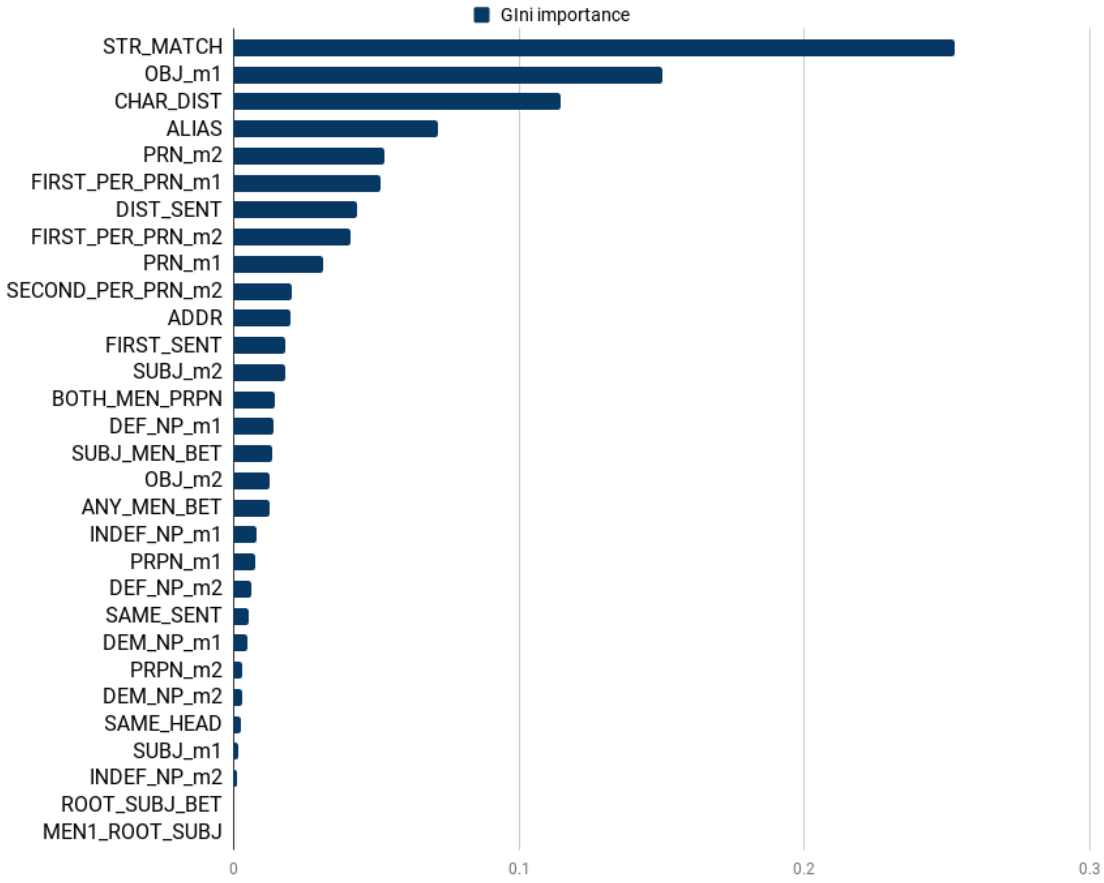


Figure 7.4: Feature importance using GI

classifier from the results in Table 7.8.

	Experiments	CoNLL Score
SVM (RBF)	CAT-I	46.05
	CAT-I + CAT-II	51.57
	CAT-I + CAT-II + CAT-III	**55.00**
Neural net	CAT-I	46.14
	CAT-I + CAT-II	54.62
	CAT-I + CAT-II + CAT-III	**57.76**

Table 7.9: Results with different feature categories

Since there are many mentions which are coreferent to each other because of their string similarity, lexical features contributes significantly to the accuracy. This accuracy is almost the same with both the classifiers. With both the classifiers, CAT-II features improves the CoNLL score by a significant number. Inspite of the few number of features present in CAT-III, the improvement it achieves is comparable to the improvement by CAT-II. The same evaluation is done with Ontonotes dataset (Pradhan et al., 2012) to see the impact of grammatical role features on a dataset with longer documents. The average size of documents in Ontonotes is ˜50 sentences, which makes the documents bigger compared to our dataset. Table 7.10 shows the results with different feature categories on Ontonotes dataset. The addition of CAT-III features including grammatical role features is not giving improvement over CAT-II with both the classifiers. This demonstrates the importance of grammatical role features to coreference resolution in shorter documents as opposed to longer documents.

7.7.4 Hybrid Approach

Here we discuss the results with hybrid approach discussed in Section 7.6. As mentioned, hybrid approach helps to segregate the classification, thus improving the selection of right knowledge sources. Table 7.11 compares the results of hybrid approach with non-hybrid approach for different classifiers discussed earlier. The non-hybrid approach results reported here the same as in Table 7.8.

Across all the classifiers hybrid approach has a better accuracy over the pure machine

	Experiments	CoNLL Score
SVM (RBF)	CAT-I	61.85
	CAT-I + CAT-II	**77.82**
	CAT-I + CAT-II + CAT-III	77.71
Neural net	CAT-I	61.85
	CAT-I + CAT-II	**78.18**
	CAT-I + CAT-II + CAT-III	77.74

Table 7.10: Results with different feature categories on Ontonotes dataset

Experiments		**MUC**			**B^3**			**CEAFe**			**CoNLL Score**
		P	R	F	P	R	F	P	R	F	
Naive Bayes	ML	37.53	37.58	37.55	50.17	53.08	51.58	39.52	50.65	44.39	44.50
	Hybrid	50.75	47.37	49.00	56.01	57.85	56.92	45.11	60.65	51.74	**52.55**
SVM (RBF)	ML	50.25	59.48	54.48	51.64	63.60	57.00	49.86	57.78	53.53	55.00
	Hybrid	52.77	61.08	56.62	53.71	64.32	58.54	50.99	60.00	55.12	**56.76**
Neural net	ML	54.50	61.69	57.87	54.30	65.27	59.28	50.45	60.38	54.97	57.37
	Hybrid	55.10	62.56	58.59	54.69	65.51	59.61	50.90	60.75	55.39	**57.87**

Table 7.11: Results comparing hybrid approach with ML based approach. P:precision, R:recall, F:F-measure.

learning based approach. Neural network classifier gives the best CoNLL score. We also evaluate the hybrid approach with gold mentions. Figure 7.5 describes the improvement of hybrid approach over pure ML across all classifiers with gold mentions.

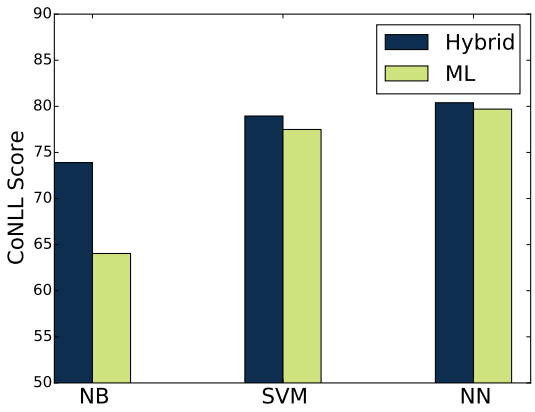


Figure 7.5: Results (CoNLL score) with gold mentions across different classifiers (NB: Naive Bayes NN: Neural network)

Categorical error analysis is done with Cort error analysis tool (Martschat et al., 2015b; Martschat and Strube, 2014). Errors are categorized by the type of the anaphoric mention (nominal, pronoun, demonstrative and proper noun). Figures 7.6 compares the recall errors between hybrid approach and pure ML approach (referred as ML in the figures) for Naive Bayes, SVM and neural network classifiers and Figure 7.7 compares precision errors. Recall errors account for the missing coreference links, whereas precision errors account for the undesired coreference links in the system output.

There is a small reduction in recall errors contributed by the proper noun anaphoric mentions with the hybrid approach, revealing the effectiveness of rule based sieve in handling mention pairs involving proper noun anaphoric mentions except for Naive Bayes classifier. But Naive Bayes improves on precision with the proper noun anaphoric mentions. The reason for the improvement with this category is that, proper noun anaphoric mentions are mostly resolved by the rule-based sieve. With this assumption, we expect the reduction of errors in this mention category to be higher. But the ML classifiers are also efficient in handling mention pairs involving both proper nouns. Here, the lexical similarity feature helps the classifier to a large extent. Errors with nominal anaphoric mentions are also more with ML. This is mainly contributed by the indefinite anaphoric mentions.

Though the hybrid approach handles the classification involving proper noun anaphoric

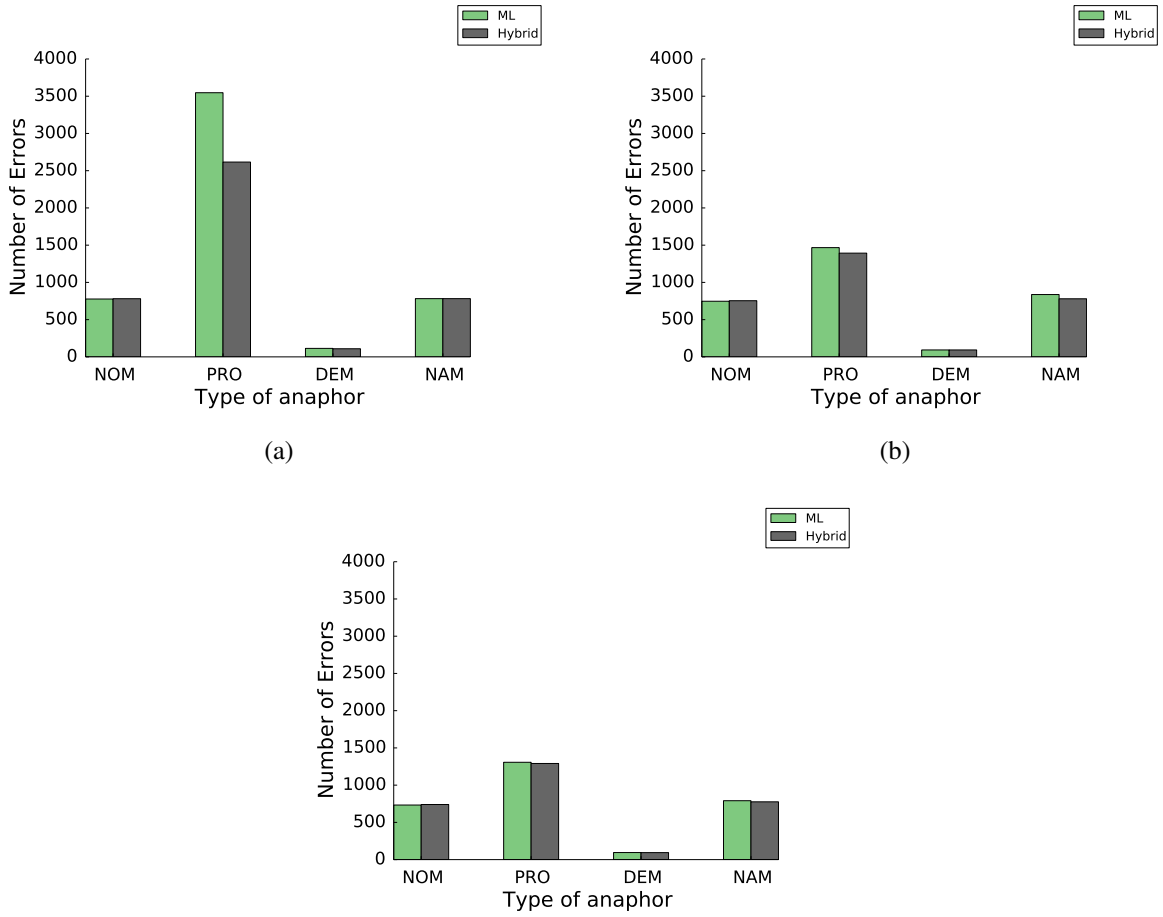
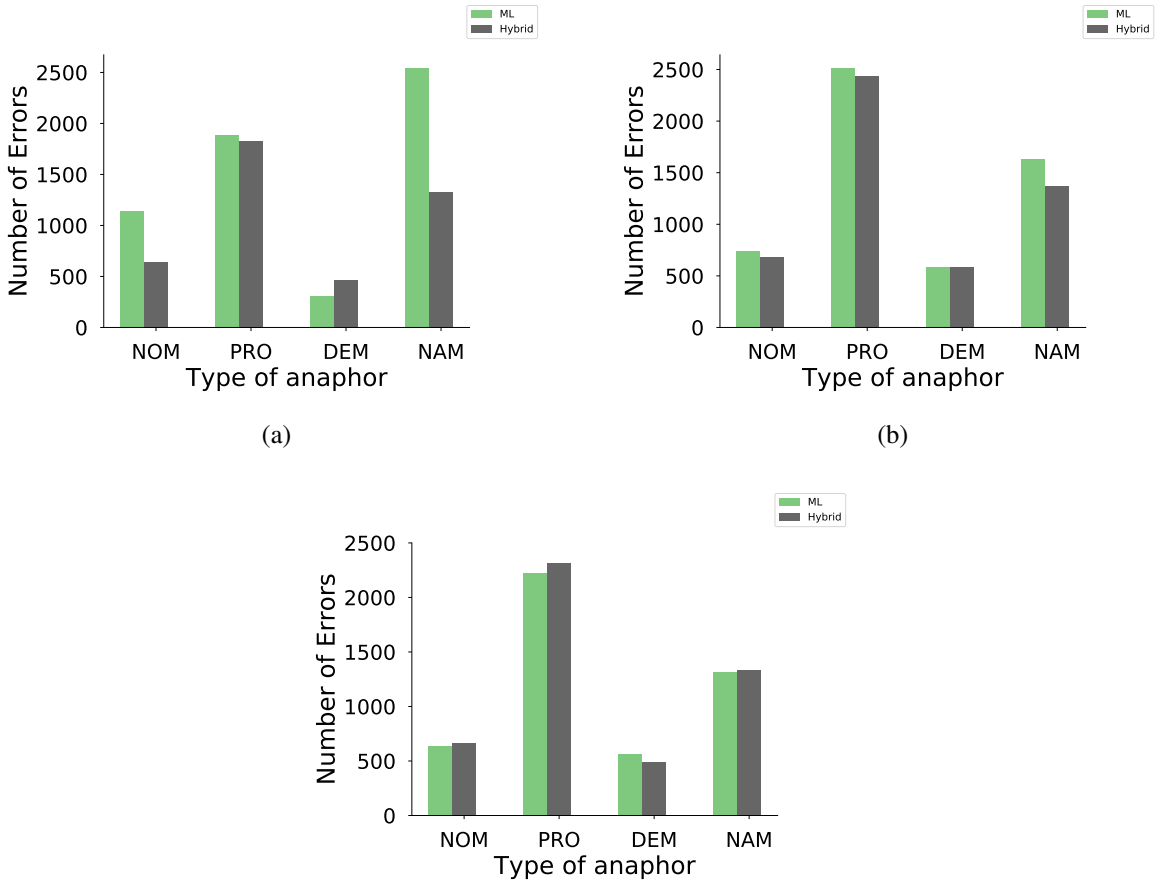


Figure 7.6: Categorized recall errors for different anaphora types with different mention-pair classifiers (NOM:nominal PRO:pronoun DEM:dem phrase NAM:proper noun) (a) Naive Bayes (b) SVM (c) Neural network



(a)

(b)

(c)

Figure 7.7: Categorized precision errors for different anaphora types with different mention-pair classifiers (NOM:nominal PRO:pronoun DEM:dem phrase NAM:proper noun) (a) Naive Bayes (b) SVM (c) Neural network

mentions better, there is a significant reduction in recall errors for pronoun category also. This is observed across classifiers. When the proper noun anaphoric mentions are handled by the rule-based sieve, the robustness of the ML classifier part of the hybrid approach improves, leading to better classification of pronoun anaphoric mentions. This is more prominent with Naive Bayes classifier, where errors due to pronoun anaphoric mentions are reduced considerably in hybrid approach. Though the major contribution for the improved accuracy in the hybrid approach comes from rule based sieve, the reduction in diversity of mention pairs handled by the ML module also contributes to the improvement.

With Naive Bayes and SVM there is a noticeable reduction in precision errors with nominal, pronoun and proper noun anaphoric mention categories, where as Neural network gives reduction in precision errors only with demonstrative anaphoric mentions. Inspite of the overall reduction in recall errors with Neural network, the increase in precision errors makes the CoNLL score improvement gained by the hybrid approach marginal with this classifier.

7.7.5 Results with an Existing Approach

Here we experiment our dataset with one of the existing coreference resolution approaches Cort (Martschat et al., 2015a), which is one among the systems those give state-of-the art results with Ontonotes dataset (Pradhan et al., 2007b).

Experiments	**MUC**			**B^3**			**CEAFe**			**CoNLL Score**
	P	R	F	P	R	F	P	R	F	
Pred	69.78	26.38	38.28	35.48	35.92	35.7	16.61	48.96	24.8	32.93
Gold	93.00	33.42	49.17	98.67	63.80	77.49	59.64	89.29	71.51	66.06
Pred§	64.96	29.48	40.56	63.80	46.92	54.07	37.29	65.38	47.49	47.37

Table 7.12: Results with Cort. Pred: with predicted mentions Gold: with gold mentions Pred§: with predicted mentions by our mention detection approach

We experiment with predicted mentions identified by Cort system itself (Pred), gold mentions (Gold) and predicted mentions identified by our mention detection approach (Pred§). The Pred§ experiment make the system independent of the problems with Cort's mention detection on our dataset. This is to show that the accuracy improvement by our approach is not just by

the virtue of our mention detection approach. For easy comparison the best performing results with our approach on predicted and gold mentions are given in Table 7.13.

Experiments	**MUC**			**B^3**			**CEAFe**			**CoNLL Score**
	P	R	F	P	R	F	P	R	F	
Pred	55.10	62.56	58.59	54.69	65.51	59.61	50.90	60.75	55.39	57.87
Gold	74.06	75.30	74.54	83.79	86.50	85.03	81.19	80.27	80.63	80.06

Table 7.13: Best performing results with our approach. Pred: with predicted mentions Gold: with gold mentions

The results show that our approach performs better on all the mention configuration. There is a major difference w.r.t predicted mentions and gold mentions. Even though Pred§ experiment get rid of the glitches in Cort's mention detection, the accuracy is inferior to our approach.

7.8 Summary

Information retrieval from music related text is an integral part of MIR augmenting audio content based MIR. We investigate the task of coreference resolution identifying its relevance to improve relation extraction from music related discussion forums. In this chapter, we introduced a coreference resolution approach for forum posts in Rasikas.org discussion forum following a supervised mention-pair model. We discussed the features for coreference resolution including the novel introductions and the modifications, considering the domain and nature of text. Bayesian network with hand-engineered network structure was found better for mention pair classification with a smaller dataset, and not with an extended dataset with sufficient data. The hybrid approach for classification gave better accuracy over pure machine learning based classification. We also discussed an approach for mention detection considering the nature of mentions in this dataset. The results with one of the state-of-the-art approaches clearly show the importance of a specific approach for coreference resolution for this dataset in specific domain and of different nature.

Chapter 8

Improved Best-First Clustering

As mentioned in the previous chapter, we follow mention-pair model for coreference resolution. Mention-pair model comprises two steps; mention pair classification and mention clustering. Clustering step in the mention-pair paradigm forms the chain of coreferent mentions from the mention pairs classified as coreferent in the classification step. Pair wise classification decisions are utilized for partitioning coreferent mentions in clustering (Ng, 2010). Best-first clustering, the widely used coreference clustering technique, takes the candidate antecedent with the highest classification confidence as the antecedent for an anaphoric mention. As most of the other coreference clustering methods, this method also considers only the classification confidence between the anaphoric mention and the candidate antecedent (Ng and Cardie, 2002b; Aone and Bennett, 1995). All our previously reported coreference resolution results are with best-first clustering. Our observations affirms the relevance of the relation between the candidate antecedents to coreference clustering. If there is a strong relation between two candidate antecedents, then one of them is more likely to be the antecedent of the anaphoric mention, compared to other candidates which have no relation between them. Based on this, we propose a modification to best-first clustering which considers the classification confidence between the candidate antecedents along with classification confidence with the anaphoric mention for clustering decision.

There are a few existing approaches for clustering. When the best-first picks the candidate antecedent with the highest classification confidence as the antecedent for an anaphoric mention, the closest-first approach selects the closest preceding coreferent mention in the discourse (Soon et al., 2001). Aggressive-merge approach selects all coreferent mentions to the anaphoric mention and make it part of the same coreferent chain (McCarthy and Lehnert, 1995). Our easy-

to-implement modification to best-first clustering intends to improve coreference resolution on Indian classical music forums. We observe a modest but statistically significant improvement over the best-first clustering for this dataset. Our discussion starts with describing best-first clustering in Section 8.1. This is followed by our approach in Section 8.2 discussing the formalized approach and the variants experimented. Section 8.3 discusses the experiments and the results of our approach and its variants, comparing them with the best results obtained so far.

8.1 Best-First Clustering

In the mention-pair model, for each anaphoric mention (m_{ana}), mention pairs are formed with the candidate antecedent mentions preceding the anaphoric mention. Mention pair classification classifies these mention pairs as coreferent or not. From the coreferent mention pairs, best-first clustering selects the mention pair having the highest classification confidence score, and takes the candidate antecedent from that mention pair as the antecedent. The probability estimate of the mention pair classification serves for the confidence score.

$$m_{ant} = \underset{m_c \in candidate\ antecedents}{argmax} P((m_c, m_{ana})) \tag{8.1}$$

Where $P((m_c, m_{ana}))$ denotes the classification probability estimate associated with the mention pair (m_c, m_{ana}).

For instance consider this sample forum post with mentions in bold.

Snehapriya *is the topic of this thread. Has this forum discussed* ***rAga snEhapriya****. There is one composition in* ***this raga*** *AFAIK,* ***kamalabhava sannuta*** *by* ***citraveeNa ravikiraN****. Is* ***this raga*** *known by another name* ***vaiShNavi*** *?*

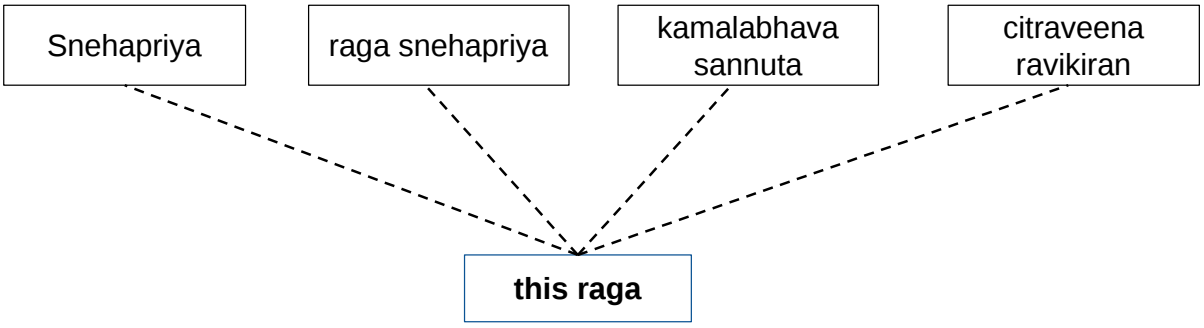


Figure 8.1: An example scenario of antecedent selection taken from a forum post

Figure 8.1 shows the anaphoric mention *this raga* in this text (last sentence) and the candidate antecedents classified as coreferent with it during mention pair classification step (dotted

line→coreference relation). In this example, if the pair associated with the candidate antecedent *kamalabhava sannuta* has the highest probability estimate value, it will be taken as the antecedent for *this raga*.

The modification to best-first clustering proposed here, modifies the confidence score associated with a mention pair (m_c, m_{ana}), based on the cues obtained from other candidate antecedents in support to this coreferent decision. Other candidate antecedents which support the coreferent relation of this mention pair are called *support* mentions.

8.2 Improved Best-First Clustering

This method is motivated by the fact that when an anaphoric mention is found coreferent with multiple candidate antecedents, the candidate mentions which are coreferent to each other are more likely to be the antecedent, compared to any other mention having no coreferent relation with other candidates. The following is an alternative description to the motivation. A chain of coreferent mentions in a document mostly has many mentions in it. If an anaphoric mention is coreferent with one among them, it is very likely that it is also coreferent with other mentions preceding it in the same chain.

Figure 8.2 shows the coreferent relations from the above discussed example. The only difference with Figure 8.1 is that it also shows coreferent relations between candidate antecedents which are considered in our method to identify the best antecedent (dotted line→coreference relation, bold line→strong coreference relation). The strong coreference relation between the candidates *Snehapriya* and *raga snehapriya* makes them better candidates over *Kamalabhava Sannuta* inspite of the highest probability estimate value associated with the the latter's classification. Here for the candidate *Snehapriya*, mention *raga snehapriya* is a support mention, making it a highly probable antecedent to *this raga*. While clustering, a candidate antecedent having a coreferent relation with other candidate antecedents of an anaphoric mention makes it a better candidate. This is the basement of the proposed modification to best-first clustering.

While best-first clustering depends solely on probability estimate associated with mention pair classification to determine confidence score, we propose to look for a method which finds the support for a candidate antecedent from other candidate antecedents and utilize this for computing confidence score along with probability estimate. Candidate antecedent having support from other candidate antecedents has better chances of getting accepted as the antecedent

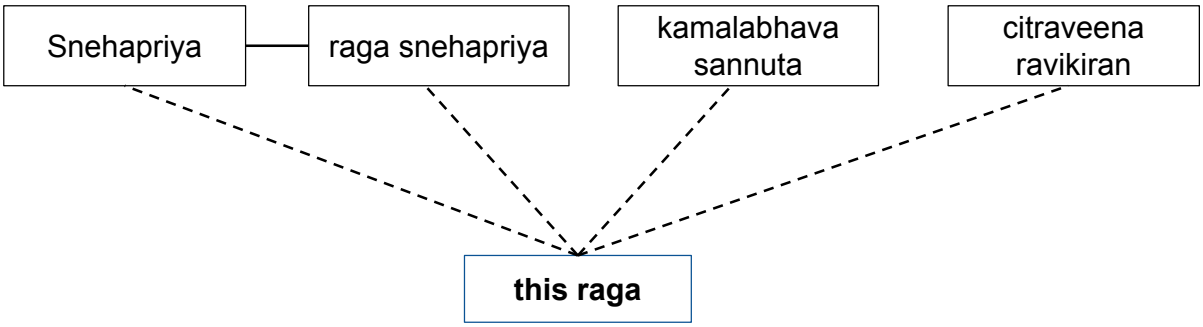


Figure 8.2: An example scenario of antecedent selection taken from a forum post

of the anaphoric mention (like *Snehapriya* in the example). The mention pair involving the candidate antecedent and support mention (another candidate antecedent) is termed as *support mention pair*. A mention is considered for support only if the classification confidence between the mention and the candidate antecedent is greater than the defined threshold ($conf\_thresh$). For *raga snehapriya* to be a support to *Snehapriya* while resolving the antecedent for *this raga*, the classification confidence of the pair (*Snehapriya*, *raga snehapriya*) has to be greater than $conf\_thresh$.

As mentioned, our mention pair classification follows a hybrid approach combining a rule-based approach with a machine learning based approach. The rule-based sieve classifies mention pairs which can be easily classified with a set of defined rules like coreference due to lexical similarity. Rest of the mention-pairs depends on machine learning based classification. Rule-based classifications are done with a higher confidence and a high confidence value (1) is assigned to these classifications as probability estimate value. Such mention pairs play a crucial role in this approach, as support decision is dependent on the classification confidence between the candidate antecedent and the support mention. In the example, the mention pair (*Snehapriya*, *raga snehapriya*) is classified by the rule-based sieve with a probability estimate value 1, making it a strong support mention pair for this case.

This clustering method identifies all such support mentions for a candidate antecedent and computes the **support score** (refer Algorithm 2). The new confidence score (non-probabilistic value) associated with a mention pair, combines the classification confidence (probability estimate) and the support score. This is computed as the linear combination of classification probability estimate and the support score associated with this mention-pair (refer Equation 8.2). This confidence score replaces the probability estimate in Equation 8.1 to find the best antecedent for an anaphoric mention.

8.2.1 Algorithm

Algorithm 2 Compute coreferent support score

Require: mention pair for which support score has to be computed($(\mathfrak{m}_{ant}, \mathfrak{m}_{ana})$), coreferent mention pairs from the document(all_mpairs), confident mention pair threshold($conf\_thresh$)

Ensure: Support score($supp$)

1: $supp \leftarrow 0$

2: $confident\_mpairs \leftarrow$ mention pairs in all_mpairs classified coreferent with prob. est. $> conf\_thresh$

3: **for all** (m_i, m_j) in all_mpairs **do**

4: **if** $(m_j == \mathfrak{m}_{ana})$ **AND** $((m_i, m_{ant}) \in confident\_mpairs$ **OR** $(m_{ant}, m_i) \in confident\_mpairs$) **then**

5: $supp \leftarrow supp + P((m_i, m_{ant}))$

Algorithm 2 describes the method to compute the support score of a candidate antecedent for an anaphoric mention (m_{ana}). The support score ($supp$) is computed for all candidate antecedents of this anaphoric mention. The method takes a mention pair involving a candidate antecedent (eg. (*Snehapriya*, *this raga*)) and all the coreferent mention pairs in the document as input. Mention pairs with a probability estimate greater than pre-defined threshold are considered for identifying the support (step 2). Step 4 defines the condition to be satisfied for a coreferent mention pair to be considered as a support mention pair for the candidate antecedent (eg. *Snehapriya*). The condition says that, the second mention of the pair must be m_{ana}. The latter part of the condition (after first AND) makes sure that m_i is coreferent with m_{ant} with classification probability estimate greater than the defined threshold ($conf\_thresh$), by checking if this pair belongs to $confident\_mpairs$. Support score ($supp$) is the sum of the classification probability estimate associated with all such support mention pairs ($P((m_i, m_{ant}))$ or $P((m_{ant}, m_i))$) . In the example, taking the candidate antecedent as *Snehapriya*, the former part of the condition assures the identified support mention is coreferent with *this raga*. *raga snehapriya* is one candidate that satisfies this. All the other 3 mentions shown in Figure 8.2 also satisfy this. Latter part checks whether *raga snehapriya* has a coreferent relation ($> conf\_thresh$) with the candidate antecedent *Snehapriya*. This is satisfied for this instance; hence *raga snehapriya* is a support mention to candidate antecedent *Snehapriya* for the anaphoric mention *this raga*.

The confidence score is now computed using

$$confidence\ score = \lambda P_e + (1 - \lambda) supp, \lambda \in (0, 1) \tag{8.2}$$

where P_e is the probability estimate associated with the mention pair classification and $supp$ is the support score associated with the mention-pair. λ decides the weightage of P_e in the confidence score.

8.2.2 Dynamic λ

Here we present a modification to the discussed clustering method. The confidence score computation is modified to have different λ values depending on the mention pair instance. This is based on the assumption, λ is directly proportional to the classification confidence associated with the mention pair. The method in Equation 8.3 takes the probability estimate value associated with the mention pair classification as its classification confidence.

$$\lambda = kP_e, k \in (0, 1) \tag{8.3}$$

where k is a constant. An alternate method is devised to decide classification confidence for the purpose of dynamic λ. Here classification confidence is computed using n different classifiers on the test data. Training data is partitioned to train these n classifiers. Testing is done on the actual test data and the variance of the classification result on a test mention pair instance is considered as its confidence of classification. Intuitively, higher variance should adversely affect classification confidence, hence λ is computed as

$$\lambda = \frac{1}{1 + clsf\_var} \tag{8.4}$$

where $clsf\_var$ is the variance of classification results from n classifiers. To maintain λ between 0 and 1, 1 is added to $clsf\_var$ in the denominator.

8.3 Experiments & Results

As discussed, our system follows the mention-pair model with a hybrid classification approach. Conventional features and the features which are found to be more important for this domain are employed (refer Section 7.3 in Chapter 7). Experiment setup is same as discussed in the

Experiments		MUC			B^3			CEAFe			CoNLL Score
		P	R	F	P	R	F	P	R	F	
SVM (RBF)	BF	52.77	61.08	56.62	53.71	64.32	58.54	50.99	60.00	55.12	56.76
	supp-BF	53.19	61.54	57.06	53.84	64.90	58.85	51.10	60.14	55.25	**57.06**
	supp-BF-1	53.17	61.52	57.04	53.84	64.88	58.85	51.10	60.13	55.25	57.05
	supp-BF-2	53.12	61.44	56.98	53.81	64.74	58.77	51.05	60.08	55.20	56.98
Neural Net	BF	55.10	62.56	58.59	54.69	65.51	59.61	50.90	60.75	55.39	57.87
	supp-BF	55.36	62.85	58.87	54.80	65.97	59.87	51.00	60.87	55.50	**58.08**
	supp-BF-1	55.37	62.85	58.87	54.80	65.94	59.86	51.00	60.87	55.50	58.08
	supp-BF-2	55.09	62.54	58.57	54.72	65.75	59.73	50.91	60.77	55.40	57.90

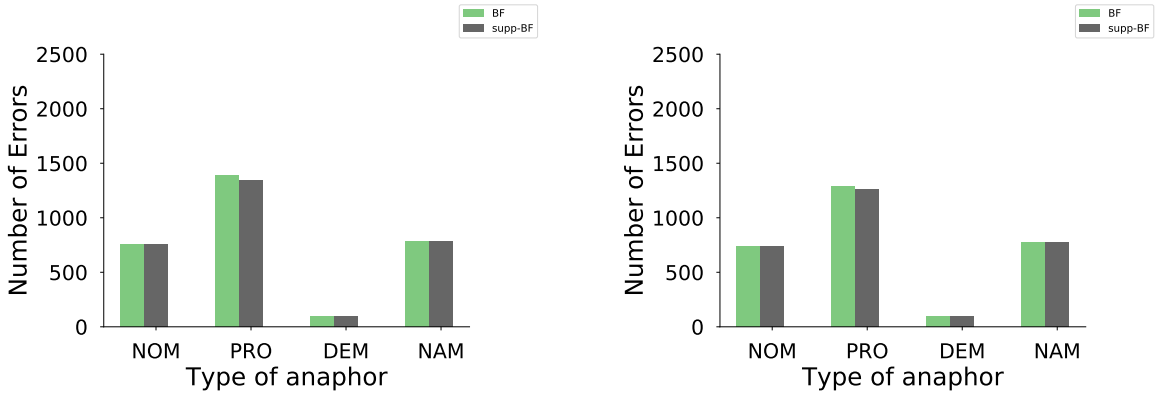
Table 8.1: Results with different classifiers (P,R,F)→ (P:Precision, R:Recall, F:F-measure), CoNLL score of significant improvements are in bold.

previous chapter. Results are reported with k-fold (5 folds) cross validation on predicted mentions. Results are averaged across different folds. The consistency of the methods is validated across 2 different classifiers which were found performing better in our previous experiments, *viz.*, Support Vector Machine with RBF kernel (SVM) and multi-layered Feed-Forward Neural Network (Neural Net). Results are reported with MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAFe (Luo, 2005) metrics. The average of F-measures from all these metrics is taken as CoNLL Score.

Table 8.1 compares the accuracy between the modifications to best-first clustering method on predicted mentions. 'BF' shows the result with best-first clustering with no modification, '*supp-BF*' with the proposed modification, '*supp-BF-1*' and '*supp-BF-2*' with the dynamic λ variations of our method. The results are reported with the best performing values for the parameters; supp-BF: λ =0.5 $conf\_thresh$=0.8 supp-BF-1: k: 0.5 $conf\_thresh$: 0.8 supp-BF-2: n classifiers =9 $conf\_thresh$: 0.8. Parameter tuning is done taking neural network as the mention-pair classifier with the development set.

With the two classifiers, supp-BF produces a noticeable improvement in accuracy compared to best-first clustering (BF). supp-BF-1 and supp-BF-2 produce no improvement over supp-BF, but gives better CoNLL score compared to the baseline BF.

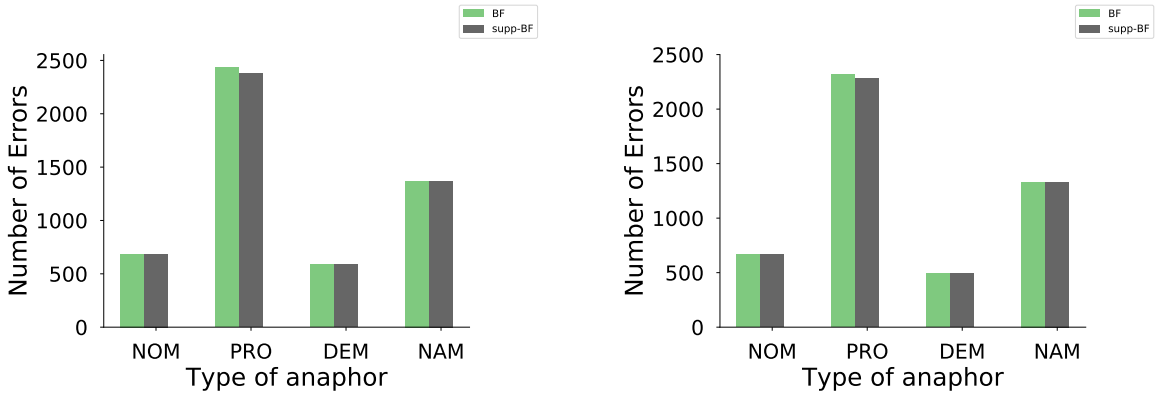
Figure 8.3 compares recall errors between supp-BF and BF experiments with both the classifiers. With both the classifiers, there is a reduction in errors with pronoun anaphoric type



(a) (b)

Figure 8.3: Categorized Recall errors for different anaphora types with different mention-pair classifiers (NOM:nominal PRO:pronoun DEM:dem phrase NAM:proper noun) ((a) SVM (b) Neural network)

and there is no difference in errors with other anaphoric types. As mention-pairs involving proper noun (NAM) anaphoric mentions are handled by the rule-based sieve with higher classification confidence, there is no improvement with supp-BF on this anaphora type. The precision errors plotted in Figure 8.4 from the same experiment also shows that the reduction in errors is only with pronoun anaphoric category. These analysis are produced with Cort error analyzer (Martschat et al., 2015b).



(a) (b)

Figure 8.4: Categorized precision errors for different anaphora types with different mention-pair classifiers (NOM:nominal PRO:pronoun DEM:dem phrase NAM:proper noun) ((a) SVM (b) Neural network)

The significance of the accuracy improvement is tested with a paired t-test on CoNLL scores (Cai and Strube, 2010). Paired t-test is relevant here, since the metrics for computing the accuracy are specific to the task of coreference resolution. Paired t-test check for the consistency in accuracy improvement across different samples from the test dataset. For this, the test set is divided into 20 sub-samples and CoNLL score is computed for each sub-sample. There is a significant improvement in CoNLL score for all the variants of our method over the baseline ($p < 0.05$) with SVM and neural network. Evaluation is also done with gold mentions of the same dataset. Here also, there is a significant improvement in accuracy with supp-BF.

8.4 Summary

While following a mention-pair model, coreference clustering is as important as mention pair classification, which makes this investigation on clustering relevant. This chapter discussed an approach that refines best-first clustering, utilizing the candidate antecedent's relation with the other candidate mentions. In a way, this approach utilizes cues from the context in discourse, rather than just depending on the candidate mentions for coreference decision. This proposed method gives better accuracy on the Rasikas.org dataset which is statistically significant, whereas the variations give improvement over the baseline, but not over the basic variant. Even though this method is motivated by the cases in our dataset, this may apply to any dataset with similar nature.

In this method, the mentions considered for finding a support for a candidate antecedent confines to other candidate antecedents. For future, we plan to explore how other mentions and words in the context can be utilized better for improved clustering.

Chapter 9

Semantic Class Identification

Understanding the semantics play a crucial role in the resolution of pronouns and nominals. Semantic class (named entity class) is a form of semantic knowledge, the identification of which is crucial for NLP problems like coreference resolution, where semantic compatibility between the entity mentions is imperative to coreference decision. To a larger extent, semantic class compatibility can help to reduce the confusions in resolving a mention. For rasikas.org dataset also, semantic class is an important feature for coreference resolution. In generic domain data, these semantic classes include person, location, organization, geo-political entity (GPE) etc. For this domain of data, it is required to introduce a new set of classes along with the conventional named entity classes for improved semantic compatibility checking. Since the term 'semantic class' appears more frequently in coreference resolution literature, we will stick on to that term instead of 'named entity class'.

The aim of this task of semantic class identification is to classify a named entity into one out of the defined semantic classes. As the case with our dataset, short and noisy text containing the entity makes it challenging to extract the semantic class of the entity through the context. We introduce a method for semantic class identification for a given entity, using the web when the entity boundaries are known. This must be distinguished from named entity recognition which involves determination of named entity boundary, followed by identification of its class. The input for semantic class identification is a string indicating a named entity, while the output is one among the semantic classes. In other words, our formulation of semantic class identification for a specific domain assumes that the entity boundaries are given. Certain domains, however, may have specific named entity classes in addition to or excluding some of these classes. For instance for information extraction from biomedical text, domain specific classes like protein,

DNA, RNA , cell are introduced (Kazama et al., 2002). In case of such specific domains, there may be several challenges: (i) There may be insufficient or no annotated data available for training a named entity recognition system (ii) The text may be noisy making it difficult to consider context, or (iii) the semantic classes include domain-specific classes.

The proposed method will be beneficial for specialized domains where data and class label challenges exist. Apart from person and location included in standard semantic classes, here we also consider raga, song, instrument and music concept. We present three approaches to use the web: (a) a baseline; rule-based approach that uses a structured web repository, (b) a supervised approach that uses search engine results and topic models, and (c) a supervised approach that improves upon (b) with task-specific hierarchy of classifiers.

The rest of the chapter is organized as follows. Section 9.1 describes the relevance of semantics to the problem of coreference resolution and discuss existing attempts to utilize semantics for coreference resolution. This section also introduces our approach and discuss the closest existing approaches in named entity recognition. Section 9.2 discusses the semantic classes related to the domain of Indian classical music. The approach and the baseline with Freebase are discussed in Section 9.3. Section 9.4 details the experiments and results on selected entities comparing the baseline and our approaches. Section 9.5 shows the results on improvement in coreference resolution accuracy when semantic class is integrated as a feature.

9.1 Semantics for Coreference Resolution

It has been identified that pragmatics and world knowledge are of prime importance to coreference resolution (Recasens, 2010; Mitkov, 1999). The challenges in integrating this to coreference resolution system is a major bottle neck to improving coreference resolution accuracy. There has been several attempts to bridge this gap. Ji et al. (2005) tries to incorporate semantic information for coreference resolution with the aid of an ontology describing relations between entities. The proposed Relational Coreference Model (RCM) takes the semantic relation between the context entities along with the participating mentions. The approach by Poesio et al. (2004) utilizes Google and Wordnet to compute the semantic distance between the mentions. Modjeska et al. (2003) devises a method to use semantic knowledge from web through lexico-syntactic patterns to resolve anaphoric mentions with 'other' or 'another' modifiers. Ponzetto and Strube (2006) investigates on features computed from semantic knowledge sources includ-

ing Wordnet (Miller, 1995) and Wikipedia along with information from semantic role labeling. Bean and Riloff (2004) and Martschat et al. (2012) also depend on Wordnet for identifying semantic class of a mention. The corpus and web based technique for analyzing semantic compatibility between mentions, checks for the likelihood of a candidate antecedent to occur in the same context of the anaphoric mention (Yang et al., 2005). With the above discussed approaches using information from external sources, it is easier to identify 'North Korea' as a country, or a mention 'the animal' corefer with another mention 'the big cat'.

Among the different semantic information useful for coreference resolution, semantic class is crucial and has been experimented in different approaches for coreference resolution as a feature. Soon et al. (2001) discusses semantic class agreement feature which classifies a mention into one of these semantic classes; *female, male, person, organization, location, date, time, money, percent* and *object*. Here semantic class identification is done through senses of a mention extracted from Wordnet. A learning based approach for classifying a noun phrase into its semantic class is discussed by Ng (2007a). They observed an improvement in the coreference resolution results with this approach, compared to heuristically determined semantic class. In another work, Ng (2007b) developed a named-entity recognizer giving good performance with MUC-6 dataset, for computing semantic class feature. Another learning based approach is discussed by Huang et al. (2009b) making use of Stanford named entity recognizer and Wordnet for semantic features.

Our method consists of three steps, and leverages on web as a knowledge repository, in order to perform the target classification. The utility of our method lies during setting up a semantic class identification system for new, specific domains. In such cases, coreference resolution task may require to have specific semantic classes for proper distinction between the entities. Also, in many cases, the context of the named entity whose class needs to be determined, may not be available. It is in such cases that the the capability of search engines and other online knowledge bases to retrieve relevant information for a named entity, is beneficial. This motivates the idea of using search engines for gathering documents for identifying the semantic class of an entity. Google has more than 30 trillion web pages indexed (statisticbrain, 2016), making it a rich source of information for any domain. The main distinction of this task as compared to named entity recognition is the absence of entity boundary detection.

Existing approaches for named entity recognition (NER) combine entity boundary identification and named entity class identification. There exists quite a large number of supervised

learning approaches for NER. Most of these approaches rely on an annotated dataset from similar domain for training the system. SVM based NER discussed in Isozaki and Kazawa (2002) classifies every word in a sentence through features related to the word and the preceding and succeeding words. This system is trained with CRL (Communication Research Laboratory) data prepared for IREX (Information Retrieval and Extraction Exercise (Sekine and Eriguchi, 2000)). MUC-6 and MUC-7 dataset served for training in the HMM based approach in Zhou and Su (2002), which used word features, semantic features and gazetteer based features. CRF based method in McCallum and Li (2003) used CoNLL-2003 English shared task data for training.

Web is used as a resource in some of the researches for NER and NER related tasks. The unsupervised approach for an NER related task in Etzioni et al. (2005) describes a web based approach to bootstrap for identifying more candidates in particular classes, given some seed candidates as input. Whitelaw et al. (2008) proposed an approach to perform named entity recognition on entire web through a supervised approach. A bootstrap based method is employed in this approach to generate training data from the web. The unsupervised approach discussed in Nadeau et al. (2006) generates a large gazetteer list from web and this is then used during disambiguating and classifying entities in a given document using simple heuristics, taking context of each entity into account. Similar approaches use web resources like Wikipedia for building an extensive gazetteer list for NER (Ratinov and Roth, 2009). Karaa (2011) proposed a method to gather training data from web with the learning examples for each class. The major distinction of our approach with the existing approaches is the utilization of web instead of the context of an entity while finding the named entity class of the entity.

9.2 Semantic Classes in Indian Classical Music Domain

In our dataset with forum posts from Rasikas.org, the discussion content is related to Carnatic music. Therefore there are entities present which are specific to this domain. Unlike data from general English domain, this domain-specific dataset requires the entities to be classified into domain-specific classes along with conventional named entity classes for better resolution of coreference. Consider the following forum post.

Sri Ragam$_{raga}$ is the asampoorna mela equivalent$_{music_concept}$ of K Priya$_{raga}$ acc to MD's$_{person}$ school. Thyagaraja$_{person}$ gave life

to K.Priya$_{raga}$ with his excellent compos, where as MD$_{person}$ never touched this raga. In Sri ragam$_{raga}$ we have plenty of compos by the trinity incl the famous Endaro$_{song}$. Sri Ranjani$_{raga}$ is a lovely janya of K Priya$_{raga}$ with plenty of compos by both T$_{person}$ & MD$_{person}$.

In this post instance, the noun phrases highlighted in yellow are the entities. The semantic class of the entities are subscripted. Here, knowing the semantic class of the mentions in the text, make it easier to resolve the anaphoric mentions *his* and *this raga*. Provided the semantic class information, *his* can only be associated with *Thyagaraja* or *MD*, and *this raga* with *K Priya* or *Sri Ragam*.

Considering the content of the forum posts, we decided to have the following semantic classes *viz.* person, raga, song, music instrument (hereafter 'instrument'), music concept (hereafter 'concept'). Table 9.1 shows some instances of each class from the dataset.

Class	**Examples**
Person	*Sri Tyagaraja Swami, M. S. Subbulakshmi*
Raga	*Mayamalavagowla, Surabhi*
Song	*dEvAdi dAva sadAshiva, Isha paahimaam*
Instrument	*Veena, Mridangam*
Concept	*Arohana, Janya*

Table 9.1: Named entity classes and examples

As mentioned, forum posts have noisy content in the form of a few grammatical errors, less structuring and spelling discrepancies. Spelling discrepancies are found more with named entities where the entities are spelled variably in different posts. For example *'Muthuswami Dikshitar', 'Dikshithar'* and *'diksitar'* refer to the same person.

The context of occurrences of certain entities have nothing much to tell about the class of the entity. Also, the context can be very similar for classes like song, raga, concept. In the following example, it is difficult to infer *balahamsa* as a raga and *mysore vAsudEvacharya* as a person.

w.r.t balahamsa, IMO it seems to be characterised (nowadays at least) mainly by variations of one prayoha - r/mgs ?

I heard a recording of the mysOre vAsudEvacharya krithi "mahAtmulE teliyalEru" sung by SK Vaagesh (probably from an AIR program) at a friend's place few years back.

Also, the extensive usage of Indian terms in text makes it harder to infer class from the context. There are instances where an entity appears alone as a separate sentence. This usually happens with composition names, followed by description in the subsequent sentences. The afore-mentioned characteristics of this dataset affirm the need for depending on the web for semantic class identification.

9.3 Devising a Web-Based Mechanism & Setting Up Classification Mechanism

Web serves as a general knowledge repository, that can be effectively harnessed for the task at hand. This particularly holds true in case of specific domains such as ours, where general-purpose knowledge repositories may not contain the required information.

We present three approaches for named entity class identification.

9.3.1 Baseline: Heuristic-Based Approach That Uses Freebase

Freebase is a vast repository of world knowledge extracted from popular wikis and stored as a database of structured knowledge (Bollacker et al., 2008). It is rich with information from specific domains like Indian classical music. Figure 9.1 shows information on Carnatic composer Thyagaraja on Freebase. This motivates the first heuristic-based approach for identifying the semantic class with the help of certain information fields in Freebase database [1].

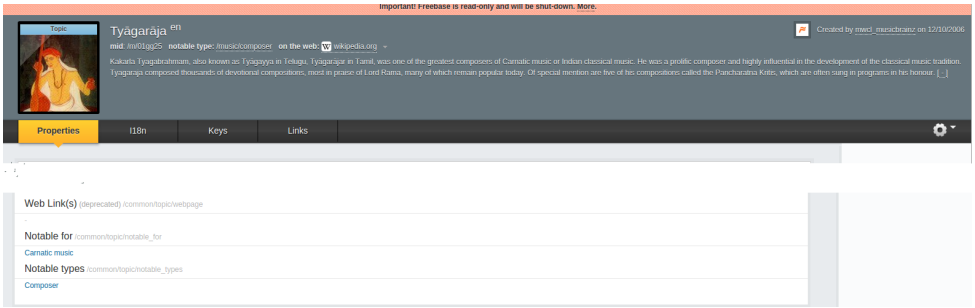


Figure 9.1: A Freebase page on Carnatic composer Thyagaraja (Relevant portion of the page)

[1] Freebase.com was officially shut down on 2 May 2016.

Algorithm 3 Semantic class identification through Freebase (approach 1)

1: **procedure** NE_CLASSID_FREEBASE($entity$) ▷ NE class of entity

2: Get Google suggestions for $entity$

3: $sel\_suggestion \leftarrow \underset{gs \in suggestions}{\text{argmax}} \; sim(entity, gs)$ ▷ gs with highest similarity with entity

4: Search $sel\_suggestion$ in Freebase to identify the type of $entity$

5: If the Freebase entity does not have type information is_a pattern is searched in Freebase description to know the type **return** semantic class

This approach is described in Algorithm 3. We try to minimize spelling discrepancies using *Google Suggest* and among the suggestions, we consider the suggestion with the highest similarity with input entity string for the subsequent steps. The robustness of *Google Suggest* in handling spelling discrepancies is put to use here to obtain better search results. Jaro-Winkler distance (Winkler, 1999), a type of string edit distance is employed to get the similarity, capable of giving more importance to the initial part of the entity words. The similarities at the beginning of the string are more significant than the ones towards the end. The entity strings from the forum are less likely to have spelling discrepancies at the initial part.

The selected string from *Google Suggest* suggestions is searched in the Freebase. The type (/common/topic/notable types) of the selected entity obtained from Freebase is taken as the semantic class of the input entity. This type is mapped to one of the semantic classes defined for this domain, except for a few which are considered as 'other'. If type is not present for an entity, we search for '*is a*' pattern in the description available with Freebase. Mostly raga entities are identified through this pattern in the description.

9.3.2 Supervised Classification Based on Web Search

This method relies on documents returned by the search engine for identifying the semantic class of an entity. Given the fact that Bing/Google has a large number of indexed pages, the chances of getting relevant pages for an entity even from a narrow domain is quite high. A classifier is pre-trained for classifying documents returned by the search engine to the relevant semantic class. Algorithm 4 describes the procedure for training the model and classification of an entity string.

The classifier is trained with handpicked documents for each semantic class. The doc-

Algorithm 4 Semantic class identification through web search (approach 2)

1: **procedure** SEMANTIC_CLASSID_WEB_TRAIN
2: Get documents for each semantic class
3: Get word clusters using LDA-based topic models from all the documents
4: Train the bag of words classifier with the document set for each semantic class
5: **return** Learned model

6: **procedure** SEMANTIC_CLASSID_WEB_TEST($entity, learned\_model$)
7: Get top k web search results for $entity$ string
8: Get the web content of the top k results
9: Combine the retrieved content into single document and then classify with the $learned\_model$
10: **return** semantic class

uments for person, raga and instrument classes come mostly from Wikipedia[2], whereas song and music concept related documents are from other sources. The classifier uses bag-of-words model for document classification. We use probabilistic models based on LDA (Blei et al., 2003) to discover clusters of words called topics. Using LDA we get the top 10 topic words for each class of documents. All the associated words with these topics collectively forms the feature vector (bag of words). These topics represent the themes underlying in the dataset. To avoid named entities getting into the bag of words, all the proper nouns in the text are masked before applying LDA.

The semantic class identification procedure gets the top k web search results for the searched entity string. The main web content of these k results are extracted[3]. The content extracted from these websites are merged to form a single document. This document is classified as one of the semantic classes with the pre-trained model (output of semantic_classid_web_train).

9.3.3 Hierarchical Hybrid Classification Based on Web Search

As the third approach, we consider a hierarchical classification approach. In this case, we segregate classification of concepts and songs using a rule-based method. In this approach, the learned supervised classifier will classify only the entities which are not classified by the rule-

[2]https://en.wikipedia.org

[3]Python library Boiler Pipe is used for main html content extraction.

based classifiers for song and concept. The method is depicted in Figure 9.2. An input entity string is given to song classification module to identify the entity as song or not. A few different heuristics are tried for song classification. One method checks for if the majority of the web search results are links to music websites. The exhaustive list of 143 music websites is used to check for if a returned link is a music website or not. A simplified version of this method is tried to check if the first link returned by Google search is a music website or not.

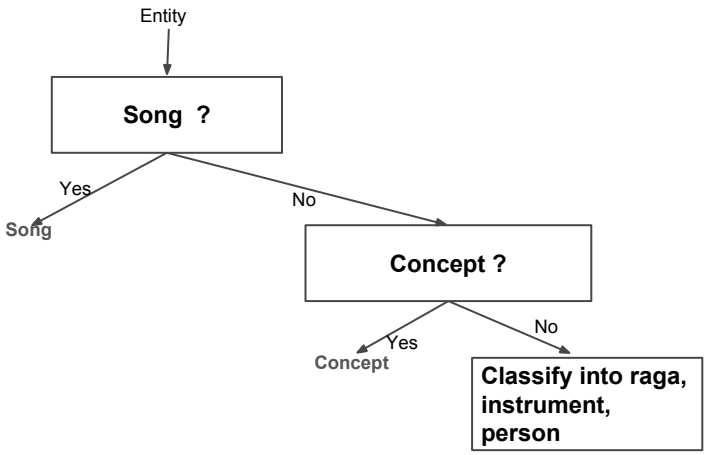


Figure 9.2: Separation of concept and song classification from the rest using hierarchical classification.

The entities which are classified as not song by the song classifier are passed on to the concept classifier. The concept classifier follows a gazetteer based approach with a gazetteer covering most of the concepts in Carnatic music. The entities having a Jaro-Winkler distance based similarity above a defined threshold value are classified as concepts. The entities which are not classified as concepts are passed on to the document classifier for getting classified as one among person, raga and instrument classes. There are 138 music concepts in the concepts gazetteer.

9.4 Experiments and Results

9.4.1 Experiment Setup

We consider 5-class classification for our experiments. Our classes are: person, raga, song, instrument, and concept. Since there is only a few occurrences of location class instances, we do not consider location as a label.

9.4.2 Comparison of Methods

Table 9.2 shows the performance of the baseline heuristic-based method. Out of 619 test entities, this method assigns no semantic class to 254 entities. The reported result takes into account only the entities classified by the method. Considering the classified entities, the overall precision is 0.77, recall 0.43 and F-score 0.55.

Class	**Precision**	**Recall**	**F1**	**Support**
Concept	1.00	0.01	0.02	87
Instrument	0.83	1.00	0.91	5
Person	0.72	0.77	0.74	118
Raga	0.81	0.39	0.52	124
Song	0.19	0.42	0.26	31
avg / total	**0.77**	**0.43**	**0.55**	**365**

Table 9.2: Results of Freebase based identification

In addition, the confusion matrix for this method is shown in Table 9.3. We see that even among the entities for which semantic class is identified, the mis-classification is high. Concept instances are getting mis-classified always, since a meaningful type could not be seen in Freebase corresponding to any concept instance. The 'other' class mentioned in the confusion matrix includes the instances which are classified to types (ex. film) which cannot be mapped to the defined classes.

	concept	**instrument**	**person**	**raga**	**song**	**other**
concept	1	36	17	11	22	0
instrument	0	0	0	0	0	5
person	0	21	91	0	5	1
raga	0	32	15	48	29	0
song	0	14	4	0	13	0

Table 9.3: Confusion matrix: Freebase based identification

The supervised approach described in algorithm 4 depends on a pre-trained model for classifying an entity. Table 9.4 describes the training documents selected for each class to train the

pre-trained model. Documents for person, raga, instrument and concept are mostly taken from Wikipedia whereas, documents for the song class are handpicked from other websites. While searching, the word 'Carnatic' (our entities are from the domain of Carnatic music) is appended to the search string for improved disambiguation. For these experiments, the top-5 (k=5) web search results are taken for an entity searched. Two popular search engines *Google* and *Bing* are used for searching the entity string. The results of the method is given in Experiment A of table 9.5.

Class	**#Documents**	**#Words**
Person	150	121777
Raga	141	53969
Song	102	84722
Instrument	121	51747
Concept	50	55440

Table 9.4: Training documents

The drop in accuracy is majorly due to the confusions with concept and song classes. The confusion matrix in table 9.6 shows that the music concept and song classes are getting heavily confused with person and raga classes. The accuracy for person, raga and instrument classes is high compared to concept and song classes. Large overlap of the words in webpages related to music concept and raga classes is the primary reason for the confusions between them. The songs are getting mostly classified as person and raga. The web content available on searching a song name are media links or pages having lyrics, notation and other information related to the song which is not helpful in classification to song class. The content of these websites have raga and singer information as well tending to classify them as raga or person.

Experiment B in table 9.5 shows the results of the third approach that uses hierarchical classification. Though the concept and song instances which are missed by the respective classifiers lead to low accuracy, the results are better compared to approach 2. Confusion matrix in 9.7 shows that, though the mis-classifications of concept class instances remain almost the same, the improvement in song classification leads to better overall accuracy.

	Experiment A			Experiment B			
Class	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
Concept	0.75	0.05	0.09	0.75	0.05	0.09	182
Instrument	0.86	0.86	0.86	0.86	0.86	0.86	7
Person	0.64	0.92	0.75	0.71	0.90	0.80	145
Raga	0.52	0.94	0.67	0.54	0.92	0.68	211
Song	0.62	0.11	0.18	0.86	0.65	0.74	74
avg / total	**0.63**	**0.57**	**0.60**	**0.68**	**0.63**	**0.65**	**619**

Table 9.5: Results of web search based identification

	concept	**instrument**	**person**	**raga**	**song**
concept	9	1	34	136	2
instrument	0	6	1	0	0
person	0	0	133	11	1
raga	1	0	10	198	2
song	2	0	31	33	8

Table 9.6: Confusion matrix: web search based identification (Experiment A)

	concept	**instrument**	**person**	**raga**	**song**
concept	9	1	31	141	0
instrument	0	6	1	0	0
person	1	0	131	10	3
raga	1	0	10	195	5
song	1	0	11	14	48

Table 9.7: Confusion matrix: web search based identification (Experiment B)

9.4.3 Error Analysis

In this section we analyze the errors of our method that uses hierarchical classification.

Person: Few person instances are confused with raga. For example singer 'Ilayaraja' when searched after appending word 'carnatic' returns irrelevant pages having only a few information about this composer. This may be happening because Ilayaraja has more contributions to Indian popular music compared to Carnatic. In the case of singer 'Rajalakshmi', pages having her songs get retrieved with a fair occurrences of the term 'raga'.

Raga: Many raga names are confused with person and song. The raga names having ambiguity with person names or other entities are likely to get classified as person. 'Snehapriya', 'K Priya' (short form of Karaharapriya), 'Ranjani' which are likely to be confused with Indian person names are classified as person. The search results of certain raga names return mostly links to music websites causing the song classification to classify them as song.

Song: Song names not meeting the song classification criteria tend to get categorized as one of the other classes. Song names like 'Bhairavi krithi', 'Thyagaraja krithis' having a raga name or a person name as a part of it are not classified as song. Song names for which search engine return websites with lyrics are also not classified as song.

Concept: The gazetteer based approach fails to identify many concepts which are combination of other concepts as in 'madhyama sruti', 'shuddha rishabam', 'raga alapan'. There also exists many Indian terms related to music but not music concepts like 'shishya', 'bhakti rasa', 'Kelvi gnanam' marked as concepts in the ground truth. These terms are not classified as concepts. Also, the absence of many concepts in the gazetteer is another reason for the poor performance.

9.5 Coreference Results with Semantic Class Feature

This section discusses the results after adding semantic class feature to the existing set of features for coreference resolution on Rasikas.org dataset. For all the entities in the corpus, we identified the semantic class using the hierarchical classification method described in Section 9.3.3. Table 9.8 shows the results of coreference resolution with the best performing classifiers *viz.* SVM (RBF) and neural network (Neural Net). Here we compare the results after adding semantic class feature to the best results obtained so far. With Neural Net classifier the improvement in CoNLL score is not significant, whereas there is a significant improvement with SVM

(RBF).

Experiments		MUC			B^3			CEAFe			CoNLL Score
		P	R	F	P	R	F	P	R	F	
SVM (RBF)	no-SC	53.19	61.54	57.06	53.84	64.90	58.85	51.10	60.14	55.25	57.06
	SC	54.67	62.44	58.30	53.86	65.63	59.17	50.73	60.15	55.04	**57.50**
Neural Net	no-SC	55.36	62.85	58.87	54.80	65.97	59.87	51.00	60.87	55.50	58.08
	SC	56.09	63.50	59.56	54.41	66.19	59.72	50.74	60.63	55.25	**58.18**

Table 9.8: Results with different classifiers (P,R,F)→ (P:Precision, R:Recall, F:F-measure). no-SC: without semantic class feature, SC: with semantic class feature

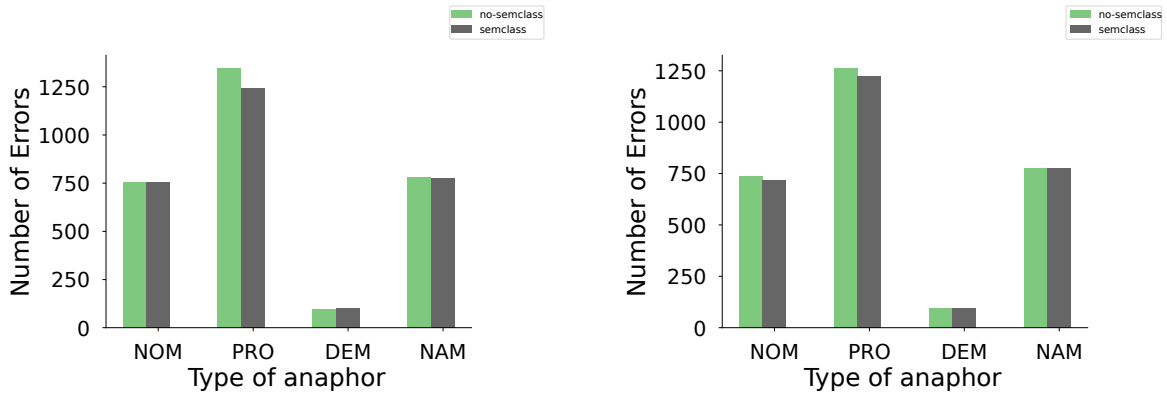


Figure 9.3: Categorized recall errors for different anaphora types with different mention-pair classifiers (NOM:nominal PRO:pronoun DEM:dem phrase NAM:proper noun) (a) SVM (b) Neural network

Figure 9.3 compares the categorized recall errors for different anaphora types between; with and without using semantic class feature for the above mentioned classifiers. With the semantic class feature, there is a reduction in errors with pronoun anaphoric mentions for both the classifiers. With SVM, semantic class feature causes a negligible increase in errors with nominal and demonstrative anaphoric mentions. With Neural Net, there is a slight decrease in errors with nominal anaphoric mention and no change is observed with demonstrative and proper noun anaphoric mentions.

Figure 9.4 shows the precision errors for the same experiments. With both the classifiers,

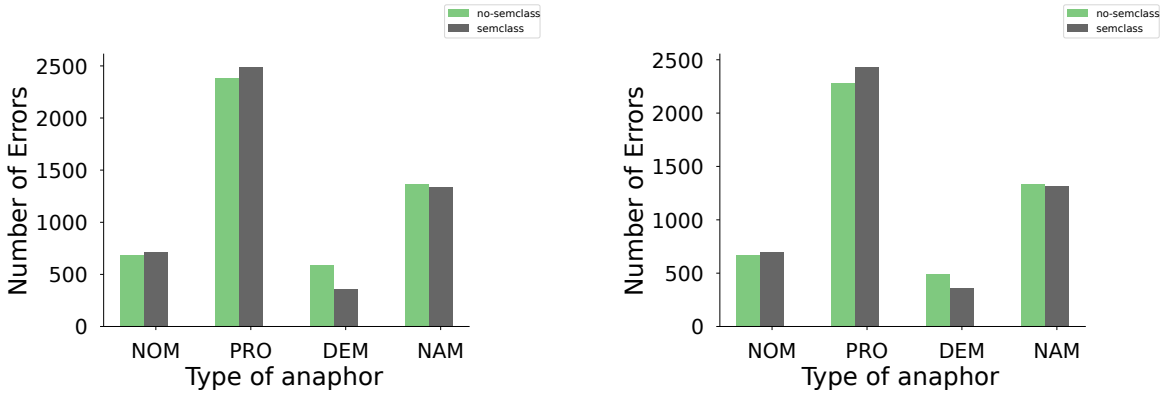


Figure 9.4: Categorized precision errors for different anaphora types with different mention-pair classifiers (NOM:nominal PRO:pronoun DEM:dem phrase NAM:proper noun) (a) SVM (b) Neural network

there is an increase in precision errors with nominal and pronoun anaphoric mentions and a decrease in errors with demonstrative and proper noun anaphoric mentions. Here, increase in pronoun category errors is quite prominent. Semantic class misclassification of certain mentions is one major reason. Since we do not use hand-labeled semantic class information for training, this mis-classification affects both coreference training and testing.

9.6 Summary

This chapter investigated methods for identifying semantic class of an entity for the coreference resolution in forum posts on Indian classical music. Such domains may be challenging due to lack of data, or presence of specific class labels. We experimented with the Rasikas.org dataset, which is utilized for our previously reported coreference resolution experiments. Our method helps to setup a platform for semantic class identification of entities making use of web resources, ignoring the context of the entities. Our methods utilizing the popular search engines to procure context are compared against the baseline approach with Freebase. The domain specificities pertaining to named entity classes are major determinants in designing the hierarchical classification model. From our case study with Indian classical music forums, it is evident that a system design driven by domain understanding is helpful. Compared to baseline approach based on Freebase, search engine based approach yields better accuracy. The

segregation of certain classes through hierarchical classification further improved the accuracy.

We also experimented semantic class as a feature for coreference resolution with our dataset. Although not significant, this helped in improving coreference resolution. Despite the importance of semantic class feature, the semantic class mis-classifications lead to marginal accuracy improvement.

For improving the semantic class identification, the method which extracts web content needs improvement to diligently filter the text to contain meaningful content related to the searched entity. Usage of nuanced LDA-based topic models will help to identify better word clusters, in the future.

Chapter 10

Novel Frontiers in Coreference Resolution

Natural Language Processing (NLP) systems intend to capture human language processing capabilities, but not always successful in achieving reasonable performance with many tasks. Coreference resolution is one among them, involving pragmatics and world knowledge along with syntactic and lexical knowledge. Coreference resolution is a discourse level problem; inferring co-referents often requires mining cues of the whole discourse. However, existing NLP systems for coreference resolution, unlike their human counterparts, are quite inefficient in extracting contextual cues from discourse. This is probably why they perform far worse than human annotators. There exist many rule based (Kennedy and Boguraev, 1996; Mitkov, 1998; Raghunathan et al., 2010) and machine learning based (Soon et al., 2001; Ng and Cardie, 2002b; Rahman and Ng, 2011c) approaches to coreference resolution. Comparing the performance of different existing systems on a standard dataset, *Ontonotes*, released for CoNLL-2012 shared task (Pradhan et al., 2012), it is quite evident that the recent systems do not have much improvement in accuracy over the earlier systems (Björkelund and Farkas, 2012; Durrett and Klein, 2013; Björkelund and Kuhn, 2014; Martschat et al., 2015a; Clark and Manning, 2015). This hints us about a saturation barrier the current paradigm may be heading towards.

In order to bridge the gap between human and machine coreference-resolution techniques, it is necessary to understand and emulate the cognitive processes underlying human coreference resolution. This chapter discusses our initiatives towards this. The first approach uses gaze information from human annotators to prune mention pairs in a coreference resolution system following a mention-pair model. This study is done with a subset of dataset taken from MUC-6 coreference dataset (Grishman and Sundheim, 1996). The second one investigates on the utility of memory networks to coreference resolution. The comprehension capabilities of

memory networks, proven with question-answering tasks, motivates this approach. Similar to the experiments by Weston et al. (2014) with memory networks on synthetic dataset for question answering, we also experiment with a synthetic dataset crafted for coreference resolution. The choice of the dataset is driven by the objective of the experiment, and for the same reason Rasikas.org dataset is not taken for these experiments.

This chapter discusses both the approaches in two different sections; Section 10.1 on eye-tracking and Section 10.2 on memory networks. The discussion on eye-tracking for coreference starts with the motivation and existing studies in psycholinguistics utilizing eye gaze information. Section 10.1.1 describes the eye-tracking experiments to collect eye gaze information from participants and Section 10.1.2 discusses our analysis on the collected eye-tracking data. Our approach to using eye-tracking information for automatic coreference resolution is discussed in Section 10.1.3, followed by experiments and results in Section 10.1.4. The discussion on memory networks starts with the motivation and existing approaches to use semantic information from the text content for coreference resolution. Section 10.2.1 describes the memory networks architecture and the modifications we experimented with. Section 10.2.3 discusses the synthetic dataset, experiments and the results. This section also compares the results with a state-of-the-art system; Cort on the same dataset.

10.1 Eye-tracking for Coreference Resolution

When devices and methods to extract human cognitive information become more accessible and handy, utilizing cognitive information for assisting many NLP tasks become far more realistic. Devising methods to utilize cues obtained from human cognitive processes in solving these tasks may help in boosting the accuracy of the NLP systems. Here the focus is on utilizing cognitive information obtained from the eye movements behavior of annotators for automatic coreference resolution. In this work, we investigate on how eye movement data obtained from readers can be helpful to the task of coreference resolution.

In this early phase of this initiative, to get rid of the noise while capturing cognitive information pertaining to coreference resolution from reading experiments, the readers for the experiments were asked to do coreference annotation. This will help to capture cognitive information relevant to coreference resolution reducing the noise. This work aims at utilizing cognitive information obtained from the eye movements behavior of annotators for automatic coreference

resolution. Here we utilize this cognitive information to improve mention-pair model, a well-known supervised coreference resolution paradigm. We first record eye-movement behavior of multiple annotators resolving coreferences in 22 documents selected from MUC dataset. The choice of this dataset is to avoid the bias a domain-specific dataset may induce. By inspecting the gaze-regression profiles of our participants, we observe how regressive saccades account for selection of potential antecedents for a certain anaphoric mention. Based on this observation, we then propose a heuristic to utilize gaze data to prune mention pairs in mention-pair model, a popular paradigm for automatic coreference resolution. Consistent improvement in accuracy across several classifiers is observed with our heuristic, demonstrating why cognitive data can be useful for a difficult task like coreference resolution.

Eye-tracking technology has been quite effective in the field of psycholinguistics to study language comprehension (Rayner and Sereno, 1994), lexical (Rayner and Duffy, 1986) and syntactic processing (von der Malsburg and Vasishth, 2011). Recently, eye-tracking studies have been conducted for various language processing tasks like Sentiment Analysis, Translation and Word Sense Disambiguation. Joshi et al. (2014) develop a method to measure the sentiment annotation complexity using cognitive evidence from eye-tracking. Mishra et al. (2013) measure complexity in text to be translated based on gaze input of translators which is used to label training data. Joshi et al. (2013) studied the cognitive aspects of Word Sense Disambiguation (WSD) through eye-tracking.

Eye-tracking studies have also been conducted for the task of coreference resolution. Cunnings et al. (2014) check for whether the syntax or discourse representation has better role in pronoun interpretation. Arnold et al. (2000) examine the effect of gender information and accessibility to pronoun interpretation. Vonk (1984) studies the fixation patterns on pronoun and associated verb phrases to explain comprehension of pronouns.

We perform yet another eye-tracking study to understand certain facets of human process involved in coreference resolution that eventually can help automatic coreference resolution. Our participants are given a set of documents to perform coreference annotation and the eye movements during the exercise are recorded. Eye-movement patterns are characterized by two basic attributes: (1) Fixations, corresponding to a longer stay of gaze on a visual object (like characters, words etc. in text) (2) Saccades, corresponding to the transition of eyes between two fixations. Moreover, a saccade is called a *Regressive Saccade* or simply, *Regression* if it represents a phenomenon of going back to a pre-visited segment. While analyzing these

attributes in our dataset, we observe a correlation between the *total regression count* and the complexity of a mention being resolved. Additionally, *mention regression count*, *i.e.,* the count of a previous mention getting visited while resolving for an anaphoric mention, proves to be a measure of relevance of that particular mention as antecedent to the anaphoric mention.

10.1.1 Creation of Eye-movement Database

We prepared a set of 22 short documents, each having less than 10 sentences. These were selected from the MUC-6 dataset[1]. Discourse size is restricted in order to make the task simpler for the participants and to reduce eye movements error caused due to scrolling.

The documents are annotated by 14 participants. Out of them, 12 of them are graduate/post-graduate students with science and engineering background in the age group of 20-30 years, with English as the primary language of academic instruction. The rest 2 are expert linguists and they belong to the age group of 47-50. To ensure that they possess good English proficiency, a small English comprehension test is carried out before the start of the experiment. Once they clear the comprehension test, they are given a set of instructions beforehand and are advised to seek clarifications before they proceed further. The instructions mention the nature of the task, annotation input method, and necessity of head movement minimization during the experiment.

The task given to the participants is to read one document at a time, and assign ids to mentions that are already marked in the document. Each id corresponding to a certain mention has to be unique, such that all the coreferent mentions in a single coreference chain are assigned with the same id. During the annotation, eye movements data of the participants (in terms of fixations, saccades and pupil-size) are tracked using an SR-Research Eyelink-1000 Plus eye-tracker (monocular mode with sampling rate of 500 Hz). The eye-tracking device is calibrated at the start of each reading session. Participants are allowed to take breaks between two reading sessions, to prevent fatigue over time.

We observe that the average annotation accuracy in terms of CoNLL-score ranges between **70.75%-86.81%**. Annotation error, we believe, could be attributed to: (a) Lack of patience/attention while reading, (b) Issues related to text comprehension and understanding, and (c) Confusion/indecisiveness caused due to lack of context.

[1] `http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html`

10.1.2 Analysis of Eye-regression Profiles

The cognitive activity involved in resolving coreferences is reflected in the eye movements of the participants, especially in the movements to the previously visited words/phrases in the document, termed as *regressive saccades* or simply, *regressions*. Regression count refers to the number of times the participant has revisited a candidate antecedent mention while resolving a particular anaphoric mention. This is extracted from the eye movement events between the first gaze of the anaphoric mention under consideration and the annotation event of this mention (when participants annotate the mention with a coreferent id).

Figure 10.1 shows the mention position (for a given mention id) in terms of the order of the mention in the document against count of regression going out from each mention to the previous mentions. The regression count for a particular mention is averaged over all the participants. As we see, average regression count tends to increase with increase in mention id, except for some mentions which may not have required visiting to the previous mentions for resolving them. The complexity of the content in MUC-6 dataset makes the spread of the regression counts dispersed. We also observe that, towards the end of the document, participants tend to regress more to the earlier sections because of limited working memory (Calvo, 2001). This increases the number of regressions performed from mentions appearing towards the end of the document.

It is worth noting that intra-sentential mentions that have antecedents within the same sentence (as in '*Prime Minister* ***Brian Mulroney*** *and* ***his*** *cabinet have been briefed today*') do not generally elicit regressions. We believe, intra-sentential resolutions are connected to processing of syntactic constraints in an organized manner, as explained by the binding theory (Chomsky, 1982). Though the number of intra-sentential mentions in our dataset is low, it is evident from figure 10.1, that they do not account for many regressions.

This above analysis on regression counts supports our hypothesis that the mentions that are regressed to more frequently have a better say in resolving an anaphoric mention.

10.1.3 Leveraging Cognitive Information for Automatic Coreference Resolution

We experiment with a supervised system following a mention-pair model (Soon et al., 2001), injecting the eye-movement information into it. Eye tracking information is utilized in the

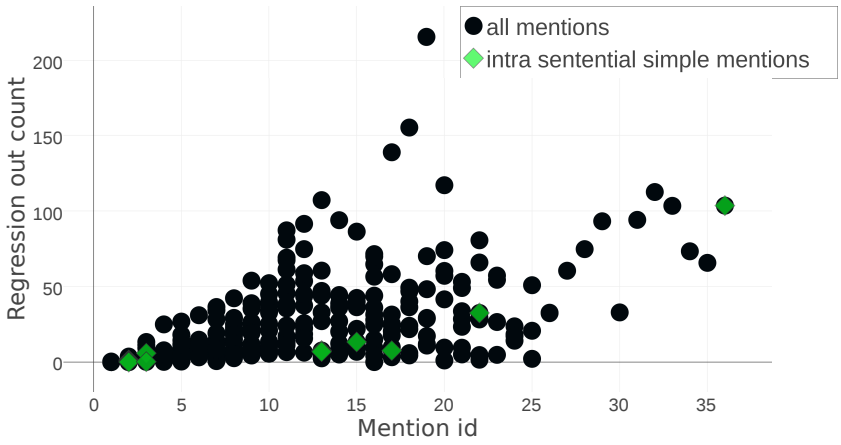


Figure 10.1: MUC-6 dataset: Mention id Vs Regression count

process of mention pair pruning prior to mention pair classification.

For Mention-pair Pruning

Given an anaphoric mention, the probability of each previous mention being selected as antecedent is computed as follows. Transitions done by a participant to potential antecedent mentions, while resolving an anaphoric mention, are first obtained from the regression profile. From this, we filter out the regressions to a candidate antecedent mention that happen between two events- (a) first fixation lands on the anaphoric mention and (b) the anaphoric mention gets annotated with an id.

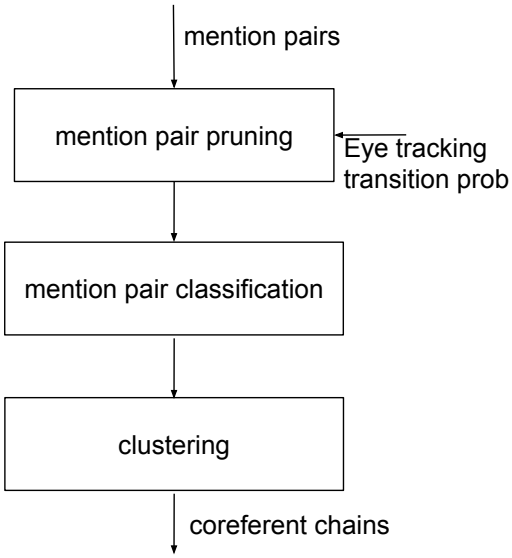


Figure 10.2: Mention-pair pruning

These regression counts from all the participants are aggregated to compute the transition probability values, as follows:

$$P_{m_i,m_j} = \frac{count(transitions\ m_j \rightarrow m_i)}{\sum_k count(transitions\ m_j \rightarrow m_k)} \tag{10.1}$$

In equation 10.1, P_{m_i,m_j} gives the transition probability value for an anaphoric mention m_j to a candidate antecedent mention m_i. count() computes the aggregated regression count over all participants. Denominator part computes for all candidate antecedents (k) of the anaphoric mention.

Transition probability thus computed for candidate mention pairs, are utilized prior to mention pair classification, filtering out irrelevant mention pairs. In the mention pair model, a mention pair (m_{ant}, m_{ana}) is formed between an anaphoric mention (m_{ana}) and a candidate antecedent mention (m_{ant}). For an anaphoric mention, the threshold probability value is computed from the number of potential candidate antecedents. $P_{thresh} = \frac{1}{\#candidate\ antecedents}$. Mentions pairs having probability less than P_{thresh} are pruned.

10.1.4 Experiments and Results

Eye-movement data driven mention pair pruning, as discussed above, is experimented across different classifiers, *viz.*, Support Vector Machine (SVM), Naive Bayes, and Multi-layered Feed-Forward Neural Network (Neural Net). We use libsvm[2] for SVM implementation and Scikit-Learn[3] for Naive Bayes implementation. The neural network classifier having an input layer, a hidden layer and an output layer is implemented using Keras[4]. For training, we consider a subset of English section of OntoNotes (v5.0) data (Pradhan et al., 2012) with 1634 documents. Testing is done with the 22 documents taken from MUC-6 dataset.

Since the main aspect of our work is mention pair pruning, we first check the mention pair pruning accuracy. We find that mention pair pruning has a precision of **87.24%**. Pruning errors may be attributed to increased number of regressions happening to mentions towards the end of the documents (refer section 10.1.2).

Performance of the system is evaluated using MUC, B^3 and CEAFe metrics. CoNLL score is computed as the average of F-measures of all the mentioned metrics. Table 10.2 shows the results across different classifiers with and without mention pair pruning. Considering the

[2]https://www.csie.ntu.edu.tw/ cjlin/libsvm/

[3]http://scikit-learn.org/

[4]http://keras.io/

CoNLL score, there is an improvement in performance across all classifiers. This improvement is contributed by the increase in precision, despite the fall in recall. Table 10.1 shows a few instances of non-coreferent antecedent-anaphora pairs which are correctly predicted as non-coreferent because of pruning.

Antecedent	**Anaphora**
here	*a treaty*
Paramount Communications Inc	*an after-tax gain of $1.2 billion*
Rogers Communications	*A Spokesman*

Table 10.1: Instances of precision errors corrected by pruning

Among all the classifiers neural network gives better accuracy, but the effective performance gain is higher with classifiers with lesser accuracy. Naive Bayes giving the least accuracy, gives the best accuracy improvement of 2.04% with mention-pair pruning. This gives the impression that systems with lower performance, are likely to benefit from the eye movement based heuristics.

Experiments		**MUC**			**B^3**			**CEAFe**			**CoNLL Score**
		P	R	F	P	R	F	P	R	F	
SVM (RBF)	unpruned	61.13	68.96	64.81	57.72	75.39	65.38	47.33	58.23	52.22	60.80
	pruned	62.67	66.99	64.76	62.62	73.71	67.71	52.00	57.83	54.76	**62.41**
SVM (Linear)	unpruned	53.33	70.93	60.88	37.64	75.02	50.13	26.56	51.27	34.99	48.67
	pruned	54.71	71.42	61.96	39.63	75.07	51.88	29.44	53.14	37.89	**50.58**
Naive Bayes	unpruned	62.85	97.53	76.44	23.23	98.03	37.56	10.53	54.22	17.64	43.88
	pruned	62.90	96.05	76.02	25.06	96.64	39.80	13.50	58.64	21.94	**45.92**
Neural Net	unpruned	64.73	71.42	67.91	63.71	77.20	69.81	52.60	61.96	56.90	64.87
	pruned	66.35	70.93	68.57	66.55	76.15	71.03	55.76	62.01	58.72	**66.11**
Berkeley coref	unpruned	84.89	58.12	69.0	84.93	47.86	61.22	82.45	37.96	51.99	60.73
	pruned	86.86	58.62	70.0	87.15	47.64	61.6	82.7	39.26	53.25	**61.61**

Table 10.2: Results with different classifiers and Berkeley coreference system with and without pruning of candidate mention pairs (P,R,F)→ (Precision, R:Recall, F:F-measure)

The performance improvement of mention pair pruning is also verified with the state of the art Berkeley Coreference Resolution system (Durrett and Klein, 2013). The choice of the

system was based on the code accessibility to make the modification required for mention pair pruning. Results of Berkeley system in table 10.2 shows that there is an improvement in CoNLL score , mainly contributed by the increase in precision.

10.2 Memory Networks for Coreference Resolution

As mentioned, this initiative is an attempt to deal with the incompetence of the existing methods for coreference resolution in comprehending the whole discourse. Comprehension of the whole discourse helps to assimilate the semantics; an important aspect for coreference resolution. There has been quite a few research in coreference resolution to bring in semantic knowledge through identification of semantic class of the entities (Ng, 2007a,b) and incorporating world knowledge with the help of sources like Wikipedia (Ponzetto and Strube, 2006; Rahman and Ng, 2011a). The semantic analysis approach for coreference resolution discussed by Hobbs (1978) takes semantics into consideration. Ng (2007b) discusses a pattern-based feature to identify corefering expressions through extracted patterns. Kehler et al. (2004) make use of predicate-argument statistics based on co-occurrence to resolve coreference. Despite these significant contributions, the achieved results show the incapability to emulate the human process of coreference resolution. The potential of memory networks (Weston et al., 2014) towards comprehending the context in a discourse motivates this initiative.

A few psycholinguistic studies on memory based processing of anaphora, investigate the processing of antecedent information from a memory representation of the discourse (Dell et al., 1983; Gernsbacher, 1989; Gerrig and McKoon, 1998; Sanford and Garrod, 1989, 2005). Experiments by Nieuwland and Martin (2016) verify the interaction between the recognition memory network and the canonical frontal-temporal language network in the human process of coreference resolution. These insights confirm the applicability of memory networks for the task.

Memory networks integrate a memory component and inference capability which are jointly used to comprehend a discourse and perform reasoning based on that (Weston et al., 2014; Sukhbaatar et al., 2015; Kumar et al., 2015). Variants of memory networks, specially designed for question answering tasks, read from the external memory multiple times before delivering the answer. Internally, they compute a representation for the input story and the question. The question representation initiates a search through the memory representation of the input and extracts relevant facts. In the subsequent step, the answer module generates the

answer based on the information got from the memory module (Sukhbaatar et al., 2015; Kumar et al., 2015). We utilize memory networks for coreference resolution, modeling it as a question answering task. The context of the mentions and its relative salience in a discourse are beneficial to resolve coreference. In practice, there are 2 ways in which coreference resolution can be assisted by memory networks, *viz.* (i) for end-to-end coreference resolution, identifying the antecedents for the anaphoric mentions (ii) for identifying the relevant sentences for resolving anaphoric mentions using attention mechanism.

End-to-end memory networks proposed by Sukhbaatar et al. (2015) for question answering is taken for our experiments. They performed question answering experiments with Facebook's synthetic dataset bAbI (Weston et al., 2015). For our experiments we create another set of synthetic data with varying difficulty levels, targeting coreference resolution. Here, each instance is a discourse and the question is on an anaphoric mention in the discourse, with the answer as its antecedent. Experiment results with memory networks on bAbI dataset is reported in terms of the accuracy of the answers whereas, our experiments also evaluate attention mechanism accuracy. We compare the prediction accuracy of memory networks with an existing state-of-the-art coreference resolution system on the same synthetic dataset. We also report results on a few modifications on memory networks.

10.2.1 Memory Networks

The end-to-end memory networks described in Sukhbaatar et al. (2015) takes input as sentences in a story ($x_1, x_2, ...x_n$), query (q) and outputs the answer (a). The sentences in the input story ($\{x_i\}$) forms the memory vectors ($\{m_i\}$), getting the word embeddings of the words within. The initial internal state u is formed from the word embeddings of the input query. The input story and the query are embedded in a continuous space through different embedding matrices (A and B), each of size $d \times V$, where V is the size of vocabulary and d is the embedding dimension.

Figure 10.3 shows the memory networks architecture with an example. The memory module has an attention mechanism responsible for identifying attention weights for each memory vector. `Softmax` over the dot product between the query representation (u_1) and each memory vector gives the probability (attention weights) associated with each memory vector w.r.t to its relevance to the given query. Attention weights are utilized to compute the weighted sum (o_1) of the memory vectors. The input query representation (u_1) is added to o_1 to obtain u_2. The above steps in the memory module are iterated depending on the number of *hops*. In each subsequent

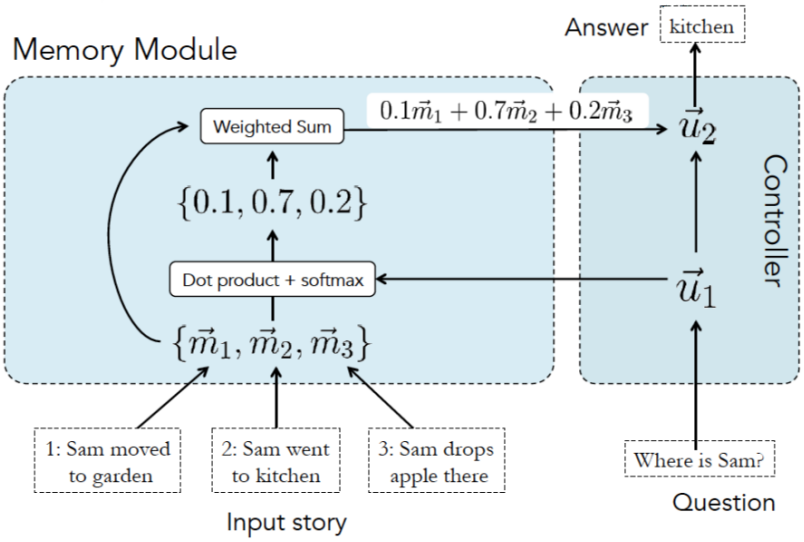


Figure 10.3: End-to-end memory networks (Weston, 2016)

iteration, u_{k+1} is computed taking u_k from the previous iteration as the input representation.

$$u_{k+1} = u_k.H + o_k \tag{10.2}$$

A linear mapping H updates u between the hops. The answer module computes $Softmax(W(o_k + u_k)$, predicting the output answer after defined number of hops.

10.2.2 Coreference Resolution as Question Answering

For our experiments with memory networks, coreference resolution is posed as a question answering problem, where the input story is the discourse containing entities and an anaphoric mention. The question is on an anaphoric mention and the answer is the antecedent entity. The following is one of the simple cases from the synthetic data.

```
Sandra went to the garden.
Mary moved to the hallway.
She is in the garden.
Who is She?    Ans:  Sandra
```

Modifications to the Network

Restricting Vocabulary: The above described memory networks architecture is designed for question answering tasks which include tasks having answers with words outside of the input

story. On the other hand, the answers in our task for coreference resolution are restricted to words within the discourse. We have introduced a modification to the answer module to switch off words outside the discourse. Our proposed modification takes a one-hot representation of the words present in a discourse. A masking layer is introduced at the output layer of the answer module. The mask vector (X_{mask}) with dimension V, has bits set for the words present in the discourse. The added layer performs element-wise multiplication between X_{mask} and the preceding output as shown in Equation 10.3 before the softmax is applied.

$$Softmax((o_k + u_k).X_{mask}) \tag{10.3}$$

Initialization of H: In the available implementation, the hidden layer matrix H in equation 10.2 is initialized with random values sampled from a normal distribution. To give uniform importance to the components in the question representation initially, this modification uniformly initialize H with ones.

`tanh` activation: As mentioned in Section 10.2.1, the probability associated with a memory vector is computed by softmax over the dot product between query representation and each memory vector. This modification applies `tanh` activation before the softmax is computed. The clipping of higher values by the `tanh` activation helps to avoid attention weights getting skewed.

While the first modification is specific to coreference resolution, the latter two are task independent.

10.2.3 Experiments

Our experiments are designed to see how memory networks can help the task of coreference resolution. All the experiments are carried out with the synthetic data.

Synthetic Dataset

Most existing memory networks based question answering research depend on synthetic dataset in order to reduce the adverse effect of noise in real-world data (Weston et al., 2015) . On similar lines, we generate 4 sets of data with different difficulty levels, keeping the vocabulary size minimal and maintaining an uniform syntactic structure. It is difficult to make valid observations with a dataset like Ontonotes (Pradhan et al., 2007a) or Rasikas.org considering the diversity in sentence structure and the vocabulary size. Each data instance has one pronominal reference

to one of the entities in the discourse. The question here is on the anaphoric mention and the answer is the antecedent mention. The 4 datasets are generated from 4 different templates randomizing the names and verbs. This synthetic data is designed in such a way that, resolution of anaphoric mentions in it requires semantic knowledge from the context. Each generated discourse has different names, actions and locations randomly picked from a pre-defined set of names, actions and locations. From the generated instances, 20% are taken for testing resulting in 11520 training instances and 2880 test instances in each dataset. Table 10.3 shows the details of dataset templates with examples.

In the all these datasets the context of the pronominal mentions serves as the main cue for resolving them. The dataset based on template 1 has 3 sentences, with 3rd sentence having a pronominal mention to one of the persons mentioned in first or second sentence. Location serves as the cue. The second template follows the same structure except for the 2 generated random sentences in order to increase the complexity of the discourse. The third template generates instances having third sentence with pronominal mention referring either to the person or the location. For each instance this is randomly chosen. Like in the second template, the fourth template has random generated sentences along with the 3rd template.

Experiment Setup

All the results are reported on the test data from the 4 synthetic datasets. One of the state-of-the-art coreference resolution systems, Cort (Martschat et al., 2015a) is chosen to compare with end-to-end memory networks (MemN2N). All the results reported with MemN2N are averaged across 10 different executions with different seeds used for training data shuffling. This is done to make the results independent of data shuffling during training. The hyper-parameters are fixed as *embedding size=20, hops=3* under the training configuration as *optimizer=Adam, #epochs=100, batch size=32, learning rate=0.01*. To make the results of Cort comparable with the answer prediction accuracy of memory networks, accuracy of Cort is computed based on the number of correctly identified coreferent mentions, instead of CoNLL score (Pradhan et al., 2012). This evaluation is valid since there is only one coreferent chain comprising 2 mentions in each synthetic dataset instance. We experiment Cort with the available pre-trained coreference model and with the model trained on training data from the corresponding synthetic dataset.

We also check for the effectiveness of attention mechanism in memory networks to aid coreference resolution, through attention mechanism accuracy. Attention mechanism accuracy

Template 1	**Template 3**
1. <name1><verb1>in the <location1>	1. <name1><verb1>in the <location1>
2. <name2><verb2>in the <location2>	2. <name2><verb2>in the <location2>
3. He/She is in <location1/location2>	3. He/She is in <loc1/loc2>. |<name1/name2>went there.
Q: Who is he/she ?	Q: Who is he/she ? |Where is there ?
e.g. Mary moved to the kitchen.	e.g. Sandra moved to the bedroom.
Daniel arrived to the office.	John went to the hallway.
She is in the kitchen.	John is playing there.
Who is She? Mary	Where is there? hallway
Template 2	**Template 4**
1. <name1><verb1>in the <location1>	1. <name1><verb1>in the <location1>
2. Random sentence 1	2. Random sentence 1
3. <name2><verb2>in the <location2>	3. <name2><verb2>in the <location2>
4. Random sentence 2	4. Random sentence 2
5. He/She is in <location1/location2>	5. He/She is in <loc1/loc2>. |<name1/name2>went there.
Q: Who is he/she ?	Q: Who is he/she ? |Where is there ?
e.g. Merlin practiced swimming.	e.g. Sita practiced wrestling.
Ram arrived to the garden.	Sandra went to the hallway.
John travelled to the office.	Ram likes swimming.
Sita likes wrestling.	Daniel travelled to the kitchen.
He is in the garden.	Sandra is working there.
Who is he? Ram	Where is there? hallway

Table 10.3: Synthetic data templates for coreference resolution

indicates, given an anaphoric mention, how capable the memory networks approach is in identifying the probable sentences to find the antecedent. The synthetic dataset has information about sentences those are relevant to the answer for each discourse instance. Attention weights obtained from memory networks are analyzed to get the sentences from the input discourse with higher attention, which in turn is used to compute attention accuracy.

10.2.4 Results

Table 10.5 compares the antecedent prediction accuracy between Cort and MemN2N. The results shows the superiority of memory networks over Cort (on both pre-trained and synthetic data trained models) in considering the context while resolving coreference. The existing feature based approaches have an inclination towards syntactical clues. Table 10.4 discusses prediction accuracy and attention accuracy with MemN2N and the modifications described in Section 10.2.2. We observe that most of the mis-predictions stem from attention errors, *i.e.* a wrong answer usually comes from a wrongly high-weighted sentence. This shows the strong dependence of the answer module on the attention mechanism.

Experiment	**Dataset 1**		**Dataset 2**		**Dataset 3**		**Dataset 4**	
	pred. acc.	att. acc.	pred. acc.	att. acc.	pred. acc.	att. acc.	pred. acc.	att. acc.
MemN2N	99.05	85.06	99.23	78.56	89.99	76.53	88.51	73.37
MASK	99.06	85.06	99.23	78.56	92.28	76.53	89.02	73.37
H-INIT	**100**	**99.83**	**100**	**99.53**	92.94	87.87	86.98	75.61
TANH	99.99	87.05	98.34	93.02	**99.75**	**92.4**	**99.55**	**89.32**

Table 10.4: Antecedent prediction accuracy (pred. acc.) and attention accuracy (att. acc.) with MemN2N and its modifications. (Accuracy in %. Best results shown in bold.)

Masking of the absent words in the discourse (MASK) has helped to improve the prediction accuracy of datasets 3 and 4. Masking helps to filter out the irrelevant words reducing the false predictions. This improvement is very intuitive, since restriction of prediction to document words is relevant to the task of coreference resolution.

The initialization of H with ones helps to reach an accuracy of 100% for datasets 1 and 2 and brings significant improvement to attention accuracy. While there is no noticeable accuracy

Experiment	DS 1	DS 2	DS 3	DS 4
Cort-pre	63.02	35.17	32.5	17.40
Cort-synth	80.42	79.90	40.66	41.04
MemN2N	99.05	99.23	89.99	88.51

Table 10.5: Comparison of antecedent prediction accuracy (%) of MemN2N with Cort. (DS: Dataset Cort-pre: results with Cort on available pre-trained model Cort-synth: results with Cort on model trained with synthetic training data)

improvement for dataset 3 and there is a reduction in accuracy for dataset 4, the improvement in attention accuracy is quite significant.

$tanh$ activation helps the system to improve the prediction accuracy significantly on datasets 3 and 4, but not for datasets 1 and 2 which have already achieved highest prediction accuracy. For datasets 1 and 2 the attention accuracy has improved compared to MEMN2N and MASK, but not compared to H-INIT. There is a considerable improvement in prediction accuracy and attention accuracy with datasets 3 and 4. $tanh$ activation enables clipping of values before `softmax` is applied, thereby preventing attention weights from getting skewed towards 0 or 1. We observed with many test instances that, when $tanh$ is not applied the memory vector with the largest attention weight in the first hop, tends to remain the largest in the subsequent hops as well. $tanh$ activation resolves this by reducing the skewness. Errors pertaining to location related pronouns with datasets 3 and 4 in the other experiments are getting reduced considerably here, resulting in improvement in accuracy.

Analysis of Cort Results

Here we explain why an existing coreference resolution approach fails to consider context based clues through analysis of distance (in terms of sentence) between the anaphoric mention and the identified antecedent in the synthetic test data. Figure 10.4 shows the distance distribution of coreferent mentions in the gold annotation for all the datasets. Sentence distances in the range 1-4 are denoted using different colors. The random sentences in DS2 and DS4 make the distance distribution broader. Figures 10.5 and 10.6 show the distance distribution of Cort output. Cort could not detect the pronominal mention 'there', making the number of coreferent distances in DS3 and DS4 less than the number of test data instances shown in the ground truth figure.

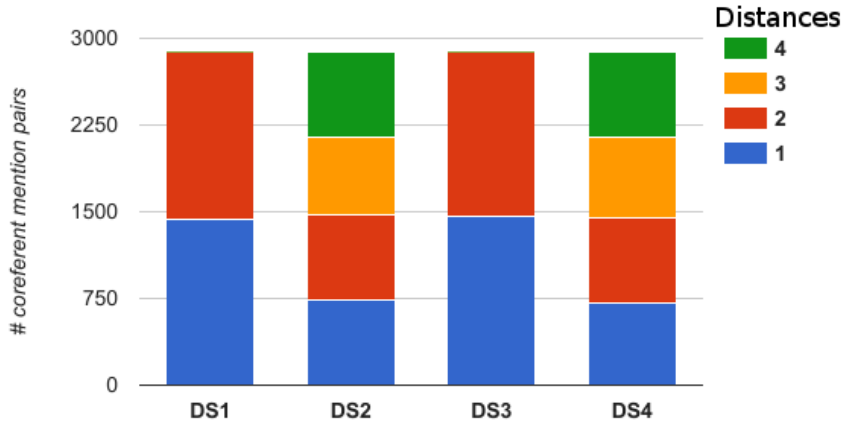


Figure 10.4: Distribution of distance between coreferent mentions in ground-truth

In a coreference resolution approach like Cort, syntactic features play a major role. Figure 10.5 shows distance distribution of coreferent mentions identified by Cort with pre-trained model trained on Ontonotes dataset. The preceding entity which forms subject in a sentence is likely to be the antecedent of an anaphoric mention in a dataset like ontonotes. When executed with pre-trained model, this leads to picking the recent subject mention as the antecedent making the distribution biased to 1. These features are designed considering the general behaviour of datasets like Ontonotes, but does not work for cases where semantic/context knowledge is important.

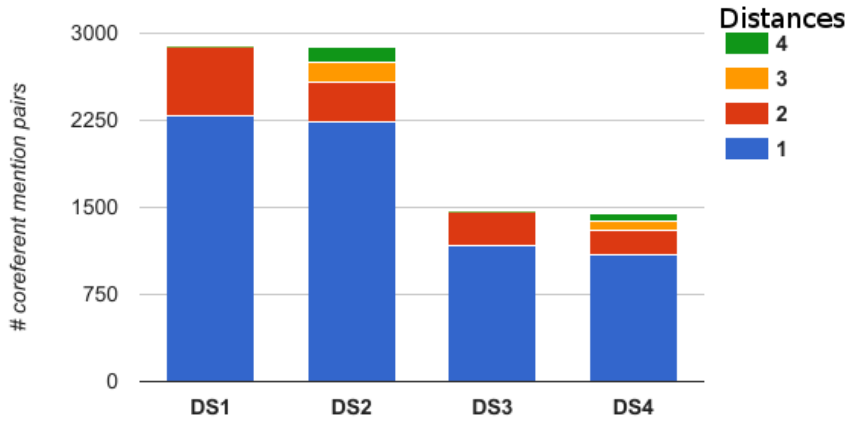


Figure 10.5: Distribution of distance between coreferent mentions identified by Cort (pre-trained model)

When trained with synthetic training set, the antecedents are not always the subject mentions in the preceding sentence based on the evidence learned from the training data. This makes the distribution of distances spread to higher distances. Even though the accuracy has improved over the experiment with pre-trained model, it is behind memory networks. From our obser-

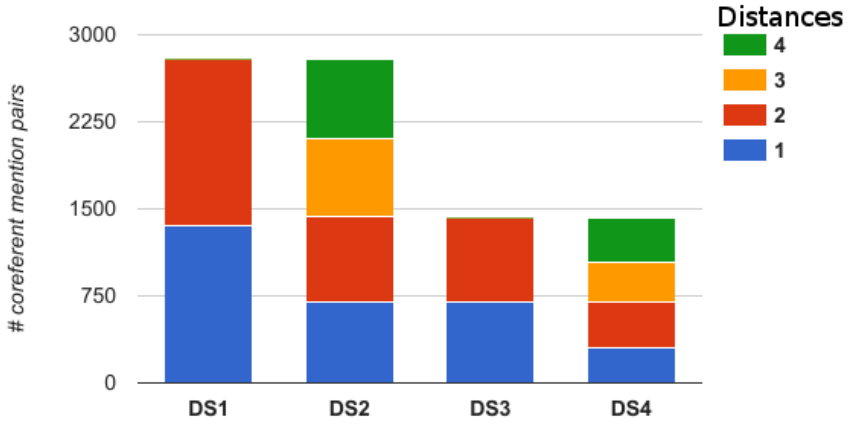


Figure 10.6: Distribution of distance between coreferent mentions identified by Cort (synthetic-trained model)

vations, we could infer that certain other features (most likely *next_token* and *preceding_token* features) in Cort take the lead role here. This makes the system to take coreference decision based on some not so relevant patterns (based on afore-mentioned features) seen in the training data, leading to inferior performance compared to memory networks.

These observations conclude that even when syntactical clues can help coreference resolution to much extent, that is not sufficient to deal with all the cases where semantic understanding is required.

10.3 Summary

This chapter discussed two different attempts to incorporate ideas from human cognition for coreference resolution, using eye-tracking input and memory networks (Weston et al., 2014). Both these attempts are motivated by the fact that the existing approaches for coreference resolution is far from the human cognition involved in coreference resolution. While the eye-tracking based approach utilized cognitive information obtained from human participants performing coreference annotation, the approach with memory networks tried to see the usefulness of the comprehending capabilities of memory networks to coreference resolution. From eye-tracking experiments, we observed that gaze transition probability derived from regression counts associated with a mention signify the candidacy of that mention as an antecedent. This helps us devise a heuristic to prune irrelevant mention pair candidates in a supervised coreference resolution approach. Our heuristic brings noticeable improvement in accuracy with different

classifiers. The experimental results comparing Cort with memory networks demonstrate the potential of memory networks. We also found that the task-driven modifications to memory networks when applied, help to achieve better prediction and attention accuracy.

The eye-tracking based approach can be further enriched to utilize eye-gaze information for (a) meaningful feature extraction for mention pair classification and (b) proposing efficient clustering mechanism. We would also like to replace our current annotation setting with a non-intrusive reading setting (say, reading text on mobile devices with camera based eye-trackers), where explicit annotations need not be required. Although the latter approach is a step towards identifying the potential of memory networks for coreference resolution, experiments are restricted to synthetic data. In the future, we propose to investigate on a memory networks based architecture on real-world data, either through attention mechanism to assist existing approaches, or through an end-to-end framework for coreference resolution.

Chapter 11

Conclusion and Future Directions

Hindustani music is popular across the world with considerable amount of recordings available in digital form. And vast amount of Hindustani music is being created and published. Construction of a knowledge base with the information extracted from this massive content body, is crucial in recommending the relevant content to the user. Along with the audio content, it is equally important to extract information from the text content pertaining to music for a comprehensive knowledge base. And this problem statement led to the research for a genre-specific approach for music information retrieval considering the requirement of the music form. This research focused on extraction of information from both audio and natural language text content. This chapter captures the findings of this research. Based on the analysis and observations discussed in the previous chapters of this thesis, the following conclusions have been derived and possible future directives have been suggested.

11.1 Conclusion

Congruent to any other structured music genre, motif detection is considered to be fundamental task in Hindustani music. Our research on motif detection was focused on melodic aspects such as detection of *mukhda* and *pakad* from Hindustani audio performances. For both the tasks, the thrust of the investigation was on finding the right time-series similarity measures and modifications to them. The title phrase or *mukhda* of a *bandish* is detected through the cue provided by the relation of a *mukhda* instance with the emphatic beat *sam*. This cue is utilized for the segmentation, prior to detection of the *mukhda* instances using similarity measures. The results shows the need of non-uniform time warping provided by DTW over the SAX similarity

measure. *Pakad* detection, an arduous task compared to *mukhda* detection, paid additional gravity on enhancing the robustness of the similarity measure which detects instances of a *pakad* from the pool of phrases. And at this juncture, relevance of learned global constraints for DTW in distinguishing instances of a *pakad* from instances of different *pakads* having high similarity in terms of phrase notation, was observed.

Detection of similarities between raga becomes critical due to its importance in finding similar Hindustani performances. We worked on extracting raga similarity from text discussion and *bandish* notations to capture similarities from two different perspectives. While certain widely accepted similarities cannot always be extracted directly from melodic content, they do appear in musicological discussions. Our approach for extracting raga similarities from text discussions aims at identifying these similarities. The modification proposed to Mikolov's approach for learning word vectors for raga words from a relatively smaller corpus, is found effective compared to the baseline. The approach to extract similarities from *bandish* notation identifies raga similarities from the basic representation of melodic content available in the form of music notation. Similarities extracted through notation serves for similarities from the dimension of melodic information; which is one of the strongest attribute from the retrieval point of view. The perplexity and clustering based evaluation validated the effectiveness of our approach to learn note-embeddings with a bi-directional LSTM based neural network.

The meta-information related to music content including artiste, details of a performance etc. are important to recommendation, because of the relevance of these attributes in determining the similarity between performances. There are many blogs, forums and other websites providing extensive information on Hindustani classical music. For our study on information extraction from text content we considered Rasikas.org, a prominent discussion forum in Carnatic music, considering the size of data available from a single source. Even though the content is related to a different genre, the nature of the text is the same and so are the challenges. Our research on coreference resolution in forum posts is driven by (i) importance of coreference resolution to improve relation extraction from the forum and (ii) the nature and domain of the text which requires a separate coreference resolution system to be developed. Following the mention-pair paradigm for supervised models, we analyzed the importance of different feature categories and implemented a hybrid approach for mention pair classification. The relevance of the introduced features and the modifications to certain features have been reflected in the coreference resolution results. We also proposed an improvement to the best-first clustering

approach, taking into account the the relationship between the candidate antecedents. This also brings in modest, but significant improvement to coreference resolution results. The approach we developed for semantic class identification utilizes web resources extracted through search, to identify the semantic class of a mention. The results indicate that this approach is effective in the identification of semantic class of mentions in any specific domain dataset where there are similar challenges. Also, the coreference results show the impact, when the semantic class is incorporated as a feature.

11.2 Future Directions

11.2.1 Motif Identification

Mid-term Goals

In our inquiry into motif detection from Hindustani music performances for *mukhda* as well as *pakad*, we considered a relatively small dataset annotated by a trained musician. Considering the scope of the study, this dataset was sufficient. We would like to extend the dataset for *mukhda* detection with performances having highly varying *mukhda* instances, which will enable to research on more robust methods. The extension of *pakad* detection dataset with performances of different ragas enabling detection of *pakads* from a diverse set of ragas will help us to explore augmented challenges in *pakad* detection.

Further work should incorporate volume and timbre dynamics so that the exploration of the constancy of raga-characteristic phrases in view of the flexibility available to performers in the improvisatory framework of Indian classical music can be more complete. Tempo is known to be a strong determiner of melodic shape, and the study of phrase intonation at high tempos is needed. Moving on from the time-series representation, a more event-based representation of the phrase in terms of basic melodic shapes is likely to be less affected by the "allowed" improvisational changes in phrase intonation. The potential of including more explicit music knowledge in motif detection can be explored in future.

So far we have not considered melodic segmentation prior to similarity computation in *pakad* detection. In future, we would like to explore methods to melodic segmentation. Cues to melodic phrase ending often, but not always, include the occurrence of a final resting note. The modeling of the perception of closure associated with phrase ending is a hitherto unexplored

problem.

Long-term Goals

The task can be modeled as a statistical classification problem, instead of classification based on thresholding the distance obtained with the similarity measure. A statistical system could capture more information about the variabilities and invariabilities within the phrases than a similarity measure can encapsulate. The available dataset does not suffice for a statistical system. We would like to explore on unsupervised or semi-supervised approaches to gather more data with limited manual effort.

11.2.2 Raga Similarity Identification

Mid-term Goals

We discussed learning representations for ragas from two diverse sources; representation in the form of word vectors learned from raga related textual discussions and representation in the form of note-embeddings from *bandish* notations. While the researched method to learn word vectors from text considers all words equally, we would like to improve the method by weighting the context words according to their importance to the domain. For the method to learn note-embeddings from *bandish* notation, our attempts to incorporate duration information were not successful. We plan to design a neural network to effectively consider duration information after looking into the necessary musicological background.

Long-term Goals

In Hindustani music, *bandish* notation is always an abstract framework for a performance. Considering the size of the available audio content, identification of similarity based on melodic attributes will be more effective if the melody representation can be extracted from the audio content. We presume that, a multi-modal approach combining *bandish* notation with the audio performance can yield better results.

11.2.3 Coreference Resolution

Mid-term Goals

- The modified Best-first clustering discussed in Chapter 8, considers only the other candidate antecedents to compute support-score for a candidate antecedent. We would like to investigate into a method utilizing the context words and other mentions in the vicinity for computing the support-score.

- The knowledge source required for resolving different types of mention pair instances are different. For instance, the relevant features for classifying a mention pair with two proper nouns are different from relevant features for classifying mention involving a proper noun and a pronoun. This motivates the scope of multi-task learning (Caruana, 1998) considering the classification of mention pairs of different types as different tasks.

- Our research on coreference resolution was wrapped up by inquiring into the potential of memory networks for coreference resolution motivated by the need of comprehending the whole document for the task. We propose to investigate on a memory networks based architecture on real-world data, either through attention mechanism to assist existing approaches, or through an end-to-end framework for coreference resolution.

Long-term Goals

The long term goal for coreference resolution is an extension to the above mentioned mid-term goal with memory networks. While the mid-term goal considers making memory networks work with real-world dataset, the same design will not guarantee the performance we aspire for. A further step is required to redesign the network to get significant improvement in coreference resolution. The existing memory networks designs are mostly motivated by the cognitive process involved in answering questions after comprehending a discourse. The cognitive process involved in coreference resolution is different from this, hence requires a different memory networks design to emulate the real cognitive process behind coreference resolution.

Bibliography

(2017). Rasikas.org. [Online; accessed 6-Jun-2017].

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA*.

Antal, P., Fannes, G., Timmerman, D., Moreau, Y., and De Moor, B. (2004). Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281.

Aone, C. and Bennett, S. W. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 122–129. Association for Computational Linguistics.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., and Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, 76(1):B13–B26.

Aucouturier, J.-J. and Sandler, M. (2002). Finding repeating patterns in acoustic musical signals: Applications for audio thumbnailing. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society.

Bagchee, S. (1998). *Nad: Understanding Raga Music*. BPI India.

Bagchee, S. (2006). *Shruti*. Rupa Publication.

Bagci, U. and Erzin, E. (2009). Inter genre similarity modelling for automatic music genre classification. *arXiv preprint arXiv:0907.3220*.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, pages 179–190.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Baldwin, B. (1997). Cogniac: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45. Association for Computational Linguistics.

Bannan, N. (2012). *Music, language, and human evolution*. Oxford University Press.

Bansal, M. and Klein, D. (2012). Coreference semantics from web features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 389–398. Association for Computational Linguistics.

Barker, C. and Pullum, G. K. (1990). A theory of command relations. *Linguistics and Philosophy*, 13(1):1–34.

Bean, D. L. and Riloff, E. (2004). Unsupervised learning of contextual role knowledge for coreference resolution. In *HLT-NAACL*, pages 297–304. Citeseer.

Belle, S., Joshi, R., and Rao, P. (2009). Raga identification by using swara intonation. *Journal of ITC Sangeet Research Academy*, 23.

Bellur, A., Ishwar, V., and Murthy, H. A. (2012a). Motivic analysis and its relevance to raga identification in carnatic music. In *Serra X, Rao P, Murthy H, Bozkurt B, editors. Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra; 2012. p. 153-157*. Universitat Pompeu Fabra.

Bellur, A., Ishwar, V., Serra, X., and Murthy, H. A. (2012b). A knowledge based signal processing approach to tonic identification in indian classical music. In *Serra X, Rao P, Murthy H, Bozkurt B, editors. Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey.* Universitat Pompeu Fabra.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bergsma, S., Lin, D., and Goebel, R. (2008). Distributional identification of non-referential pronouns. In *AcL*, volume 8, pages 10–18.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

Bhagyalekshmy, S. (1990). *Ragas in Carnatic music*. South Asia Books.

Bhattacharjee, A. and Srinivasan, N. (2011). Hindustani raga representation and identification: a transition probability based approach. *International Journal of Mind, Brain and Cognition*, 2(1-2):66–91.

Björkelund, A. and Farkas, R. (2012). Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 49–55. Association for Computational Linguistics.

Björkelund, A. and Kuhn, J. (2014). Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. *ACL, Baltimore, MD, USA, June*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Boersma, P. and Weenink, D. (1992). 2001. praat. a system for doing phonetics by computer.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Borg, I. and Groenen, P. (2003). Modern multidimensional scaling: theory and applications. *Journal of Educational Measurement*, 40(3):277–280.

Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., Versley, Y., and Zanoli, R. (2010). Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 104–107. Association for Computational Linguistics.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Buteau, C. and Mazzola, G. (2000). From contour similarity to motivic topologies. *Musicae Scientiae*, 4(2):125–149.

Cai, J. and Strube, M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 28–36. Association for Computational Linguistics.

Calvo, M. G. (2001). Working memory and inferences: Evidence from eye fixations during reading. *Memory*, 9(4-6):365–381.

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation. *Music Perception: An Interdisciplinary Journal*, 23(3):249–268.

Caruana, R. (1998). Multitask learning. In *Learning to learn*, pages 95–133. Springer.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696.

Cavnar, W. B., Trenkle, J. M., et al. (1994). N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

Celma, O. (2010). Music recommendation. In *Music Recommendation and Discovery*, pages 43–85. Springer.

Chakraborty, S., Mazzola, G., Tewari, S., and Patra, M. (2014). *Computational Musicology in Hindustani Music*. Springer.

Chakravorty, J., Mukherjee, B., and Datta, A. K. (1989). Some studies in machine recognition of ragas in indian classical music. *Journal of the Acoust. Soc. India*, 17(3&4).

Chen, B., Su, J., and Tan, C. L. (2013). Random walks down the mention graphs for event coreference resolution. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4):74.

Chen, C. and Ng, V. (2012). Combining the best of two worlds: A hybrid approach to multilingual coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 56–63. Association for Computational Linguistics.

Chen, J. and Wang, C. (2017). Investigating the role of musical genre in human perception of music stretching resistance. *arXiv preprint arXiv:1701.03274*.

Cheng, X. and Voigt, R. (2015). A deep architecture for coreference resolution. In *Proc. of the 2001 Workshop on Computational Natural Language*, pages 1–8.

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*, volume 6. MIT press.

Chordia, P. (2006). Automatic raag classification of pitch-tracked performances using pitch-class and pitch-class dyad distributions. In *ICMC*.

Chordia, P. and Rae, A. (2007). Raag recognition using pitch-class and pitch-class dyad distributions. In *ISMIR*, pages 431–436. Citeseer.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Association of Computational Linguistics (ACL)*.

Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Clarkson, P. and Robinson, T. (2001). Improved language modelling through better language model evaluation measures. *Computer Speech & Language*, 15(1):39–53.

Collins, T., Thurlow, J., Laney, R., Willis, A., and Garthwaite, P. (2010). A comparative evaluation of algorithms for discovering translational patterns in baroque keyboard works.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Connolly, D., Burger, J. D., and Day, D. S. (1997). A machine learning approach to anaphoric reference. In *New Methods in Language Processing*, pages 133–144.

Cooper, M. L. and Foote, J. (2002). Automatic music summarization via similarity analysis. In *ISMIR*.

Crawley, R. A., Stevenson, R. J., and Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 19(4):245–264.

Culotta, A., Wick, M. L., and McCallum, A. (2007). First-order probabilistic models for coreference resolution. In *HLT-NAACL*, pages 81–88.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Cunnings, I., Patterson, C., and Felser, C. (2014). Variable binding and coreference in sentence comprehension: evidence from eye movements. *Journal of Memory and Language*, 71(1):39–56.

Dagan, I. and Itai, A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 330–332. Association for Computational Linguistics.

Dakwale, P., Mujadia, V., and Sharma, D. M. (2013). A hybrid approach for anaphora resolution in hindi. In *IJCNLP*, pages 977–981.

Dannenberg, R. B. and Hu, N. (2003). Pattern discovery techniques for music audio. *Journal of New Music Research*, 32(2):153–163.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Dell, G. S., McKoon, G., and Ratcliff, R. (1983). The activation of antecedent information during the processing of anaphoric reference in reading. *Journal of Verbal Learning and Verbal Behavior*, 22(1):121–132.

Denis, P. and Baldridge, J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669. Association for Computational Linguistics.

Dighe, P., Agrawal, P., Karnick, H., Thota, S., and Raj, B. (2013). Scale independent raga identification using chromagram patterns and swara based features. In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pages 1–4. IEEE.

Ding, X. and Liu, B. (2010). Resolving object and attribute coreference in opinion mining. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 268–276. Association for Computational Linguistics.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1.

D'Souza, J. and Ng, V. (2012). Anaphora resolution in biomedical literature: a hybrid approach. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 113–122. ACM.

Durrett, G. and Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In *EMNLP*, pages 1971–1982.

Durrett, G. and Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Eck, D. and Schmidhuber, J. (2002). A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, 103.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Ehrlich, K. and Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1):75–87.

Erzin, E. et al. (2007). Automatic classification of musical genres using inter-genre similarity. *IEEE Signal Processing Letters*, 14(8):521–524.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3):131–163.

Fu, A. W.-C., Keogh, E., Lau, L. Y., Ratanamahatana, C. A., and Wong, R. C.-W. (2008). Scaling and time warping in time series querying. *The VLDB JournalThe International Journal on Very Large Data Bases*, 17(4):899–921.

Futrelle, J. and Downie, J. S. (2002). Interdisciplinary communities and research issues in music information retrieval. In *ISMIR*, volume 2, pages 215–221.

Ganguli, K. K. and Rao, P. (2017). Towards computational modeling of the ungrammatical in a raga performance.

Ge, N., Hale, J., and Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the sixth workshop on very large corpora*, volume 71, page 76.

Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition*, 32(2):99–156.

Gerrig, R. J. and McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes*, 26(2-3):67–86.

Godbole, V., Liu, W., and Togneri, R. (2015). An investigation of neural embeddings for coreference resolution. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–251. Springer.

Goldberg, Y. and Levy, O. (2014). word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.

Grober, E. H., Beardsley, W., and Caramazza, A. (1978). Parallel function strategy in pronoun assignment. *Cognition*, 6(2):117–133.

Gulati, S., Bellur, A., Salamon, J., Ishwar, V., Murthy, H. A., and Serra, X. (2014). Automatic tonic identification in indian art music: approaches and evaluation. *Journal of New Music Research*, 43(1):53–71.

Gulati, S., Salamon, J., and Serra, X. (2012). A two-stage approach for tonic identification in indian art music. In *Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra; 2012. p. 119-127*. Universitat Pompeu Fabra.

Hacioglu, K., Douglas, B., and Chen, Y. (2005). Detection of entity mentions occurring in english and chinese text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 379–386. Association for Computational Linguistics.

Haghighi, A. and Klein, D. (2010). Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Hartrumpf, S. (2001). Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7*, page 17. Association for Computational Linguistics.

Hendrickx, I. and Hoste, V. (2009). Coreference resolution on blogs and commented news. In *Anaphora Processing and Applications*, pages 43–53. Springer.

Hendrickx, I., Hoste, V., and Daelemans, W. (2007). Evaluating hybrid versus data-driven coreference resolution. In *Anaphora: Analysis, Algorithms and Applications*, pages 137–150. Springer.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4):311–338.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huang, A. and Wu, R. (2016). Deep learning for music. *arXiv preprint arXiv:1606.04930*.

Huang, S., Zhang, Y., Zhou, J., and Chen, J. (2009a). Coreference resolution using markov logic networks. *Advances in Computational Linguistics*, 41:157–168.

Huang, Z., Zeng, G., Xu, W., and Celikyilmaz, A. (2009b). Accurate semantic class classifier for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1232–1240. Association for Computational Linguistics.

Isozaki, H. and Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Ji, H., Westbrook, D., and Grishman, R. (2005). Using semantic relations to refine coreference decisions. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 17–24. Association for Computational Linguistics.

Joshi, A., Bhattacharyya, P., and Carman, M. (2016). Political issue extraction model: A novel hierarchical topic model that uses tweets by political and non-political authors. In *WASSA, Proceedings of NAACL-HLT*, pages 82–90.

Joshi, A., Mishra, A., Senthamilselvan, N., and Bhattacharyya, P. (2014). Measuring sentiment annotation complexity of text. In *ACL (2)*, pages 36–41.

Joshi, S., Kanojia, D., and Bhattacharyya, P. (2013). More than meets the eye: Study of human cognition in sense annotation. In *HLT-NAACL*, pages 733–738.

Juhász, Z. (2007). Analysis of melody roots in hungarian folk music using self-organizing maps with adaptively weighted dynamic time warping. *Applied Artificial Intelligence*, 21(1):35–55.

Junius, M., Daniélou, A., Waldschmidt, E., Waldschmidt, R., and Kaufmann, W. (1969). The ragas of northern indian music.

Karaa, W. B. A. (2011). Named entity recognition using web document corpus. *arXiv preprint arXiv:1102.5728*.

Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pages 1–8. Association for Computational Linguistics.

Kehler, A., Appelt, D. E., Taylor, L., and Simma, A. (2004). The (non) utility of predicate-argument frequencies for pronoun interpretation. In *HLT-NAACL*, volume 4, pages 289–296.

Kennedy, C. and Boguraev, B. (1996). Anaphora for everyone: pronominal anaphora resoluation without a parser. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 113–118. Association for Computational Linguistics.

Kertz, L., Kehler, A., and Elman, J. (2006). Grammatical and coherence-based factors in pronoun interpretation. In *Proceedings of the 28th annual conference of the Cognitive Science Society*, pages 1605–1610.

Kobdani, H. and Schütze, H. (2010). Sucre: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95. Association for Computational Linguistics.

Koduri, G. K. (2016). *Towards a multimodal knowledge base for Indian art music : A case study with melodic intonation*. PhD thesis, Universitat Pompeu Fabra.

Koduri, G. K. and Serra, X. (2013). A knowledge-based approach to computational analysis of melody in indian art music. In *International Workshop on Semantic Music and Media colocated with International Semantic Web Conference*, pages 1–10.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Korb, K. B. and Nicholson, A. E. (2010). *Bayesian artificial intelligence*. CRC press.

Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., Ondruska, P., Gulrajani, I., and Socher, R. (2015). Ask me anything: Dynamic memory networks for natural language processing. *arXiv preprint arXiv:1506.07285*.

Kumar, V., Pandya, H., and Jawahar, C. (2014). Identifying ragas in indian music. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 767–772. IEEE.

Kummerfeld, J. K., Bansal, M., Burkett, D., and Klein, D. (2011). Mention detection: heuristics for the ontonotes annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 102–106. Association for Computational Linguistics.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Lassalle, E. and Denis, P. (2013). Improving pairwise coreference models through feature space hierarchy learning. In *ACL 2013-Annual meeting of the Association for Computational Linguistics*.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM.

Liutkus, A., Rafii, Z., Badeau, R., Pardo, B., and Richard, G. (2012). Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 53–56. IEEE.

Logan, B. and Chu, S. (2000). Music summarization using key phrases. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages II749–II752. IEEE.

Lu, L., Wang, M., and Zhang, H.-J. (2004). Repeating pattern discovery and structure analysis from acoustic music data. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 275–282. ACM.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 135. Association for Computational Linguistics.

Martschat, S., Cai, J., Broscheit, S., Mújdricza-Maydt, E., and Strube, M. (2012). A multigraph model for coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 100–106. Association for Computational Linguistics.

Martschat, S., Claus, P., and Strube, M. (2015a). Plug latent structures and play coreference resolution. In *ACL (System Demonstrations)*, pages 61–66.

Martschat, S., Göckel, T., and Strube, M. (2015b). Analyzing and visualizing coreference resolution errors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 6–10.

Martschat, S. and Strube, M. (2014). Recall error analysis for coreference resolution. In *EMNLP*, pages 2070–2081.

McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

McCallum, A. and Wellner, B. (2005). Conditional models of identity uncertainty with application to noun coreference. In *Advances in neural information processing systems*, pages 905–912.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. *Proceedings of the International Joint Conference on Artificial Intelligence*.

McCord, M. C. (1990). Slot grammar. pages 118–145.

Meek, C. and Birmingham, W. P. (2001). Thematic extractor. In *Proceedings of the 2nd International Society for Music Information Retrieval Conference*.

Meredith, D., Lemström, K., and Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, pages 746–751.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mishra, A., Bhattacharyya, P., Carl, M., and CRITT, I. (2013). Automatically predicting sentence translation difficulty. In *ACL (2)*, pages 346–351.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 869–875. Association for Computational Linguistics.

Mitkov, R. (1999). *Anaphora resolution: the state of the art*. Citeseer.

Modjeska, N. N., Markert, K., and Nissim, M. (2003). Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 176–183. Association for Computational Linguistics.

Muller, M. (2007). *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg.

Nadeau, D., Turney, P., and Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity.

Nagy, G. (1968). State of the art in pattern recognition. *Proceedings of the IEEE*, 56(5):836–863.

Ng, V. (2007a). Semantic class induction and coreference resolution. In *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics*, page 536.

Ng, V. (2007b). Shallow semantics for coreference resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*.

Ng, V. (2009). Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 575–583. Association for Computational Linguistics.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1396–1411. Association for Computational Linguistics.

Ng, V. and Cardie, C. (2002a). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Ng, V. and Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Nicolae, C. and Nicolae, G. (2006). Bestcut: A graph algorithm for coreference resolution. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 275–283. Association for Computational Linguistics.

Nieto, O. and Farbood, M. M. (2012). Perceptual evaluation of automatically extracted musical motives. In *Proceedings of the 12th International Conference on Music Perception and Cognition*, pages 723–727.

Nieuwland, M. and Martin, A. E. (2016). A neural oscillatory signature of reference. *bioRxiv*, page 072322.

Nilsson, K. (2010). *Hybrid methods for coreference resolution in Swedish*. PhD thesis, Department of Linguistics, Stockholm University.

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

Padmasundari, G. and Murthy, H. A. (2017). Raga identification using locality sensitive hashing. In *Communications (NCC), 2017 Twenty-third National Conference on*, pages 1–6. IEEE.

Pandey, G., Mishra, C., and Ipe, P. (2003). Tansen: A system for automatic raga identification. In *IICAI*, pages 1350–1363.

Parrikar, R. (2000). parrikar.org.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Peeters, G. (2003). Deriving musical structures from signal analysis for music audio summary generation:sequence and state approach. In *International Symposium on Computer Music Modeling and Retrieval*, pages 143–166. Springer.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 143. Association for Computational Linguistics.

Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007a). Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 1(04):405–419.

Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. (2007b). Unrestricted coreference: Identifying entities and events in ontonotes. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 446–453. IEEE.

Rafii, Z. and Pardo, B. (2013). Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE transactions on audio, speech, and language processing*, 21(1):73–84.

Raghavan, P., Fosler-Lussier, E., and Lai, A. M. (2012). Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–741. Association for Computational Linguistics.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Rahaim, M. (2012). *Musicking Bodies: Gesture and Voice in Hindustani Music*. Wesleyan University Press.

Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings*

of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pages 968–977. Association for Computational Linguistics.

Rahman, A. and Ng, V. (2011a). Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 814–824. Association for Computational Linguistics.

Rahman, A. and Ng, V. (2011b). Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *Journal of Artificial Intelligence Research*, 40:469–521.

Rahman, A. and Ng, V. (2011c). Syntactic parsing for ranking-based coreference resolution. In *IJCNLP*, pages 465–473.

Raimond, Y., Abdallah, S. A., Sandler, M. B., and Giasson, F. (2007). The music ontology. In *Proceedings of the 8th International Society for Music Information Retrieval Conference; 2007; Vienna, Austria*, pages 417–422. International Society for Music Information Retrieval.

Raja, D. (2005). Hindustani music. *A Tradition in Transition, DK Printworld (P) Ltd.*

Ram, R. V. S. and Devi, S. L. (2012). Coreference resolution using tree crfs. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 285–296. Springer.

Ranjani, H., Arthi, S., and Sreenivas, T. (2011). Carnatic music analysis: Shadja, swara identification and raga verification in alapana using stochastic models. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 29–32. IEEE.

Rao, P., Ross, J. C., and Ganguli, K. K. (2013). Distinguishing raga-specific intonation of phrases with audio analysis. *ninad*, 26:64.

Rao, S. (2013). Music in motion.

Rao, S., Bor, J., van der Meer, W., and Harvey, J. (1999). *The raga guide: a survey of 74 hindustani ragas*. Nimbus Records with Rotterdam Conservatory of Music.

Rao, S. and Rao, P. (2014a). An overview of hindustani music in the context of computational musicology. *Journal of New Music Research*, 43(1):24–33.

Rao, S. and Rao, P. (2014b). An overview of hindustani music in the context of computational musicology. *Journal of New Music Research*, 43(1):24–33.

Rao, V., Gaddipati, P., and Rao, P. (2012). Signal-driven window-length adaptation for sinusoid detection in polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):342–348.

Rao, V., Gupta, C., and Rao, P. (2011). Context-aware features for singing voice detection in polyphonic music. In *International Workshop on Adaptive Multimedia Retrieval*, pages 43–57. Springer.

Rao, V. and Rao, P. (2010). Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE transactions on audio, speech, and language processing*, 18(8):2145–2154.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Rayner, K. and Sereno, S. C. (1994). Eye movements in reading: Psycholinguistic studies.

Recasens, M. (2010). *Coreference: Theory, Annotation, Resolution and Evaluation.* PhD thesis, Universitat de Barcelona.

Recasens, M. and Hovy, E. (2009). A deeper look into features for coreference resolution. In *Anaphora Processing and Applications*, pages 29–42. Springer.

Recasens, M. and Martí, M. A. (2010). Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language resources and evaluation*, 44(4):315–345.

Rich, E. and LuperFoy, S. (1988). An architecture for anaphora resolution. In *Proceedings of the second conference on Applied natural language processing*, pages 18–24. Association for Computational Linguistics.

Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine learning*, 62(1):107–136.

Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell III, J. T., and Johnson, M. (2002). Parsing the wall street journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 271–278. Association for Computational Linguistics.

Rolland, P.-Y. (1999). Discovering patterns in musical sequences. *Journal of New Music Research*, 28(4):334–350.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420.

Rösiger, I. and Riester, A. (2015). Using prosodic annotations to improve coreference resolution of spoken text. In *ACL (2)*, pages 83–88.

Sadhana (2011). Hindustani classical music- notes. [Online; accessed 28-May-2017].

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Salamon, J. and Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770.

Salamon, J., Gulati, S., and Serra, X. (2012). A multipitch approach to tonic identification in indian classical music. In *Gouyon F, Herrera P, Martins LG, Müller M. ISMIR 2012: Proceedings of the 13th International Society for Music Information Retrieval Conference; 2012 Oct 8-12; Porto, Portugal. Porto: FEUP Ediçoes; 2012*. International Society for Music Information Retrieval (ISMIR).

Sanford, A. and Garrod, S. (1989). What, when, and how?: Questions of immediacy in anaphoric reference resolution. *Language and Cognitive Processes*, 4(3-4):SI235–SI262.

Sanford, A. J. and Garrod, S. C. (2005). Memory-based approaches and beyond. *Discourse Processes*, 39(2-3):205–224.

Sekine, S. and Eriguchi, Y. (2000). Japanese named entity extraction evaluation: analysis of results. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 1106–1110. Association for Computational Linguistics.

Serra, X. (2011). A multicultural approach in music information research. In *Klapuri A, Leider C, editors. ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference; 2011 October 24-28; Miami, Florida (USA). Miami: University of Miami; 2011.* International Society for Music Information Retrieval (ISMIR).

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jorda, S., et al. (2013). Roadmap for music information research.

Shetty, S., Achary, K., and Hegde, S. (2012). Clustering of ragas based on jump sequence for automatic raga identification. In *Wireless Networks and Computational Intelligence*, pages 318–328. Springer.

Sikdar, U. K., Ekbal, A., Saha, S., Uryupina, O., and Poesio, M. (2015). Differential evolution-based feature selection technique for anaphora resolution. *Soft Computing*, 19(8):2149–2161.

Singh, S., Subramanya, A., Pereira, F., and McCallum, A. (2011). Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 793–803. Association for Computational Linguistics.

Singla, P. and Domingos, P. (2006). Entity resolution with markov logic. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 572–582. IEEE.

Sobha, L. D., Pattabhi, R. R., Ram, R., Malarkodi, C., and Akilandeswari, A. (2011). Hybrid approach for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 93–96. Association for Computational Linguistics.

Song, Y., Jiang, J., Zhao, W. X., Li, S., and Wang, H. (2012). Joint learning for coreference resolution with markov logic. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1245–1254. Association for Computational Linguistics.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., and De Ilarraza, A. D. (2012). Mention detection: First steps in the development of a basque coreference resolution system. In *KONVENS*, pages 128–136.

Sridhar, R., Subramanian, M., Lavanya, B., Malinidevi, B., and Geetha, T. (2011). Latent dirichlet allocation model for raga identification of carnatic music. *Journal of Computer Science*, 7(11):1711.

Stamborg, M., Medved, D., Exner, P., and Nugues, P. (2012). Using syntactic dependencies to solve coreferences. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 64–70. Association for Computational Linguistics.

statisticbrain (2016). statisticbrain.com.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 656–664. Association for Computational Linguistics.

Strube, M., Rapp, S., and Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 312–319. Association for Computational Linguistics.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Suja (2012). Music to my ears. [Online; accessed 27-May-2017].

Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Swarganga (2004). Swarganga.org.

Szeto, W. M. and Wong, M. H. (2006). A graph-theoretical approach for pattern matching in post-tonal music analysis. *Journal of New Music Research*, 35(4):307–321.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

TP, V., Sankagiri, S., Ganguli, K. K., and Rao, P. (2016). Structural segmentation and visualization of sitar and sarod concert audio. In *Gouyon F, Herrera P, Martins LG, Müller M. ISMIR 2012: Proceedings of the 17th International Society for Music Information Retrieval Conference; 2016; New York, USA*. International Society for Music Information Retrieval (ISMIR).

Typke, R., Wiering, F., Veltkamp, R. C., et al. (2005). A survey of music information retrieval systems. In *Ismir*, pages 153–160.

Uryupina, O. (2006). Coreference resolution with and without linguistic knowledge. In *Proceedings of LREC*, pages 893–898.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9.

van der Meer, W. (2012). *Hindustani music in the 20th century*. Springer Science & Business Media.

Venkataraman, A., Kubicki, M., Westin, C.-F., and Golland, P. (2010). Robust feature selection in resting-state fmri connectivity based on population studies. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 63–70. IEEE.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Vinutha, T. and Rao, P. (2014). Audio segmentation of hindustani music concert recordings. In *Proceedings of the International Symposium, Frontiers of Research on Speech and Music (FRSM)*.

von der Malsburg, T. and Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65(2):109–127.

Vonk, W. (1984). Eye movements during comprehension of pronouns. *Advances in Psychology*, 22:203–212.

Webber, B. L. and Joshi, A. K. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. *arXiv preprint cmp-lg/9806017*.

Weiss, R. J. and Bello, J. P. (2011). Unsupervised discovery of temporal structure in music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1240–1251.

Weissenbacher, D. and Sasaki, Y. (2013). Which factors contributes to resolving coreference chains with bayesian networks? In *14th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 200–212.

Weston, J. (2016). *ICML 2016 Tutorial on Memory Networks for Language Understanding*.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.

Whitelaw, C., Kehlenbeck, A., Petrovic, N., and Ungar, L. (2008). Web-scale named entity recognition. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 123–132. ACM.

Wick, M., Singh, S., and McCallum, A. (2012). A discriminative hierarchical model for fast coreference at large scale. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 379–388. Association for Computational Linguistics.

Wikipedia (2017a). Hindustani classical music. [Online; accessed 26-May-2017].

Wikipedia (2017b). Metre (music). [Online; accessed 29-May-2017].

Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Yang, X., Su, J., Lang, J., Tan, C. L., Liu, T., and Li, S. (2008). An entity-mention model for coreference resolution with inductive logic programming. In *ACL*, pages 843–851.

Yang, X., Su, J., and Tan, C. L. (2005). Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 165–172. Association for Computational Linguistics.

Yang, Y., Xue, N., and Anick, P. (2011). A machine learning-based coreference detection system for ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 117–121. Association for Computational Linguistics.

Zhekova, D. and Kübler, S. (2010). Ubiu: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99. Association for Computational Linguistics.

Zheng, J., Chapman, W. W., Crowley, R. S., and Savova, G. K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*, 44(6):1113–1122.

Zhou, G. and Kong, F. (2009). Global learning of noun phrase anaphoricity in coreference resolution via label propagation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 978–986. Association for Computational Linguistics.

Zhou, G. and Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics.

Publications from the thesis

Journals

1. Preeti Rao, Joe Cheri Ross, Kaustuv Kanti Ganguli, Vedhas Pandit, Vignesh Ishwar, Ashwin Bellur and Hema Murthy, *Classification of Melodic Motifs in Raga Music with Time series Matching*, Journal of New Music Research, 2014.

2. Preeti Rao, Joe Cheri Ross, and Kaustuv Kanti Ganguli, *Distinguishing Raga-specific Intonation of Phrases with Audio Analysis*, Ninad: Journal of ITC SRA, Vol. 26-27, December 2013.

Peer-reviewed Conferences

1. Joe Cheri Ross, Abhijit Mishra, Kaustuv Kanti Ganguli, Pushpak Bhattacharyya and Preeti Rao, *Identifying Raga Similarity Through embeddings learned from Compositions notation*, The 18th International Society for Music Information Retrieval Conference (ISMIR 2017), Suzhou, China, October 23, 2017.

2. Joe Cheri Ross, Rudra Murthy, Kaustuv Kanti Ganguli and Pushpak Bhattacharyya, *Identifying Raga Similarity in Hindustani Classical Music through Distributed Representation of Raga Names*, The 13th International Symposium on Computer Music Multidisciplinary Research (CMMR 2017), Porto, Portugal, September 25, 2017.

3. Joe Cheri Ross and Pushpak Bhattacharyya, *Towards Harnessing Memory Networks for Coreference Resolution*, 2nd Workshop on Representation Learning for NLP (Repl4NLP) at ACL 2017, Vancouver, Canada, August 3, 2017.

4. Joe Cheri Ross and Pushpak Bhattacharyya, *Improved Best-First Clustering for Coreference Resolution in Indian Classical Music Forums*, 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2017), Budapest, Hungary, April 17, 2017.

5. Joe Cheri Ross, Abhijit Mishra and Pushpak Bhattacharyya, *Leveraging Annotators Gaze Behaviour for Coreference Resolution*, ACL 2016 Workshop on Cognitive Aspects of Computational Language Learning (CogACLL 2016) at ACL 2016, Berlin, Germany, August 11, 2016.

6. Joe Cheri Ross, Aditya Joshi and Pushpak Bhattacharyya, *A Framework That Uses the Web for Named Entity Class Identification: Case Study for Indian Classical Music Forums*, 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2016), Konya, Turkey, April 3-9, 2016.

7. Joe Cheri Ross and Pushpak Bhattacharyya, *Coreference Resolution to Support IE from Indian Classical Music Forums*, Recent Advances in Natural Language Processing (RANLP 2015), Hissar, Bulgaria, September 5, 2015.

8. Joe Cheri Ross, Sachin Pawar and Pushpak Bhattacharyya, *Coreference Resolution for Better Information Retrieval from Indian Classical Music Forums*, 11th International Conference on Natural Language Processing (ICON 2014), Goa, India, December 18, 2014.

9. Joe Cheri Ross, TP Vinutha and Preeti Rao, *Detecting Melodic Motifs from Audio for Hindustani Classical Music*, 13th International Society for Music Information Retrieval Conference (ISMIR 2012), Porto, Portugal, October 2012.

10. Joe Cheri Ross and Preeti Rao, *Detection of Raga-Characteristic Phrases from Hindustani Classical Music Audio*, Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey, July, 2012.

Additional Collaborative Work

1. Samarth Agarwal, Aditya Joshi, Joe Cheri Ross, Pushpak Bhattacharyya and Harsha Wabgaonkar, *Are Word Embedding and Dialogue Act Class-based Features Useful for Coreference Resolution in Dialogue?*, The 15th International Conference of the Pacific Association for Computational Linguistics (PACLING 2017), Yangon, Myanmar, August 16, 2017.

Acknowledgment

I really cannot begin before looking back to my teenage days as an amatuer keyboardist at my parish church as well as my social circles, way back in the 1990s. I express deepest thanks to my parents, relatives, my well wishers and my parish who were my first audience, critics and source of inspiration.

I honestly believe that my love for music that developed over the years led me to the gates of IIT Bombay. I have no words to express my profound gratitude to this august institution and its rich history of innovation & research. Ever since I came to know about the field of music information retrieval, I never had a second thought on the area I would want to pursue my doctoral research.

Words cannot describe the gratitude I want to express to Prof. Preeti Rao and Prof. Pushpak Bhattacharyya for giving me an opportunity to do doctoral research under their guidance, and for the endless support they extended throughout my research. I am grateful to Prof. Ganesh Ramakrishnan, Prof. Saketha Nath and Prof. Om Damani for their insightful comments and directions given as part of my Research Progress Committee (RPC). My sincere thanks to Prof. Xavier Serra and Prof. Asif Ekbal for reviewing my thesis. Their comments and feedback were indispensable in having the thesis in its current form. I also appreciate the contributions from Dr. Suvarnalata Rao for the valuable discussions and helping us to validate our approaches musicologically.

My lack of knowledge in Hindustani music was comprehensively compensated by Kaustuv Kanti Ganguli, who was also instrumental in creating most of the annotated datasets for motif detection. My heartfelt thanks to this exceptionally talented musician. Special word of gratitude for Vinutha Prasad and Amruta Vidwans as their music dataset annotations initiated my research on motif detection. Big thanks to the efforts of Mrs. Jaya Jha and Rajchandra Laishram in annotating the coreference dataset.

No research is complete without constructive collaboration and cooperation from faculty

and fellow research associates. I am indebted to some of the most dignified experts in this area who were kind enough to offer support, even before I requested them to. It might be impossible to individually acknowledge their contribution in detail here, but to name a few :

Prateek Verma , Amruta, Vinutha, Kaustuv, Prateek ,Sankalp Gulati (UPF Barcelona), Gopal Krishna Koduri (UPF Barcelona), Ajay Srinivasamurthy, Mohammed Sordo, Abhijit Mishra, Aditya Joshi, Anoop Kunchukuttan, Sachin Pawar, Rudra Murthy, Vasudevan N, Kevin Patel, Girish Ponkiya, Arun Iyer, Uma Sawant , Ramkrishna Bairi, Abhijit Mishra, Sudha Bhingardive, Kiran Sawant, Arun Iyer, Swetha Srinivasan, Uma Sawant , Ramkrishna Bairi, Deepak Gupta, Darsh Shah, Samarth Agarwal, Ashish Kulkarni

– thank you from the bottom of my heart.

My internship with IBM Research was an opportunity to try a few different approaches. Thanks to Sachindra Joshi and Dinesh Raghu from IBM Research for this opportunity and the guidance. My gratitude to Prof. Mitesh who facilitated this internship opportunity.

I am equally obliged to all the mental, moral and spiritual support I received from all corners of my life. The spiritual support extended at tough times by Prof. C D Sebastian; guidance provided by Prof. Sabu Thomas (pro vice chancellor, MG University); motivation offered by Prof. Vinod Pathari, Prof. Madhukumar S D and Prof. Muralikrishnan from NIT Calicut, Prof. Soney George from Amal Jyothi College of Engineering ; care and encouragement from my cousins Davis Kuriakose, Tess Zacharias and friends like Satyanand Abraham, Kiran Irimpan, Alex Chollakal, Stephen George, Anoop George, Biju Antony, Ajay Nair, Aravind B ,Jose Joseph all these were momentous to my healthy continuance in research. I thank you all and wish you the very best. My association with the church choir of Kanjurmarg West parish was a source of energy to me. I thank all the parishioners, especially Mr. Joseph for the consideration.

The support of my family is indispensable for the successful completion of my PhD. I am indebted to my parents for being with me in the decision of pursuing PhD and supporting me. I hereby acknowledge the contribution of my sisters in this journey, as early day fans of my music. Thanks to my parent-in-laws for their constant prayer support.

Merry, my life partner, and Immanuel, my son I am indebted for life, for all the hardships you both had to go through during my days of research. If not for your patient understanding, this would never have happened. I am especially grateful to my son Immanuel for improving my art of storytelling.

Last but not the least, thanks to the God Almighty - from whom all good things including

music originates - for leading me through this wonderful experience of learning.