Coreference Resolution to Support IE from Indian Classical Music Forums

Joe Cheri Ross **Pushpak Bhattacharyya**
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{joe,pb}@cse.iitb.ac.in

Abstract

Efficient music information retrieval (MIR) require to have meta information about music along with content based information in the knowledge base. Discussion forums on music are rich sources of information gathered from a wider audience. Taking into consideration the nature of text in these web resources, the yield of relation extraction is quite dependent on resolving the entity references in the document. Among the few music forums dealing with Indian classical music, *rasikas.org* (rasikas, 2015) having rich information about artistes, raga and other music concepts is taken for our study. The forum posts generally contain anaphoric references to the main topic of the thread or any other entity in the discourse. In this paper we focus on coreference resolution for short discourse noisy text like that of forum posts. Since grammatical roles capture relation between mentions in a discourse, those features extracted from dependency parsing are widely explored along with semantic compatibility feature. On investigation of issues, the need for integrating known dependencies between features emerged. A Bayesian network with predefined network structure is evaluated, since a Bayesian belief network enacts a probabilistic rule based system. To the extent possible the superior behaviour of Bayesian network over SVM is analysed.

1 Introduction

Information extraction from music repositories involves analysis of music audio. Efficient extraction of music information require meta-data along with content based information. The need for metadata led to information extraction from blogs and forums related to music. This should contain information about artistes, performances, music concepts etc. Apart from the available literature about Indian classical music, there are a few forums and blogs having rich metadata. Extracting information from these sources help to augment music ontology for Indian classical music with meta information along with content based information. Among the two main divisions in Indian classical music, Carnatic music community is more involved in web based discussions and information dissemination. *Rasikas.org* (rasikas, 2015) is one among the prominent discussion forums where they have discussions pertaining to Carnatic music topics comprising ragas, talas, artistes etc.

Extracting information from unstructured noisy text in websites of this kind is quite challenging. Efficient extraction of relations also require resolution of entities in the documents. Apart from resolving the entities with the real world entities, the intra-relations between the entities within the discourse have to be resolved. Identification of entities is a critical step in information extraction followed by identification of relations between them.

Posts in most 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. Effecient extraction of relation is dependent on finding the exact antecedent of pronominal and nominal mentions, when it refers to another entity. It is commonly observed that the main topic of a post is referred by pronominal or alias mention. Following is a post from the forum. 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

Proceedings of Recent Advances in Natural Language Processing, pages 91–96,
Hissar, Bulgaria, Sep 7–9 2015.

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.

The presence of a large number of such sentences containing potential relations present, make coreference resolution unavoidable for information extraction from these forums. The process of checking whether two expressions are coreferent to each other is termed as coreference resolution (Soon et al., 2001). The well-known discussion forum on Carnatic music *Rasikas.org*, is taken for our study. Enrolled with a good number of music loving users, the forum discusses many relevant topics on Carnatic music providing valued information. Sordo et al. evaluated information extraction from the same forum using contextual information (Sordo et al., 2012). Integration of natural language processing methods yields better coverage for the extracted relations. Largely the entities are mentioned using pronominal and nominal mentions in this forum. Resolution of these coreferences is crucial in increasing recall of relation extraction from forums. Coreference resolution identifies the real world entity, an expression is referring to (Cherry and Bergsma, 2005).

Though a widely researched area, coreference resolution will have to be applied differently considering the characteristics of the text in these forums. Forum posts are generally short discourse of text where the entities mentioned are limited to the scope of a few sentences. Supervised approach has been widely used in coreference resolution (Rahman and Ng, 2009; Soon et al., 2001; Aone and Bennett, 1995; McCarthy and Lehnert, 1995). We examine the commonly used conventional features and its variants that suits this domain of text. Soon et al. and Vincent et al. have investigated an exhaustive list of features for coreference resolution. Most of these methods model this problem as classification of mention-pair as coreferent or non-coreferent. Research on coreference resolution for similar domains of text are reported. Ding et al. has discussed features for supervised approach to coreference resolution for opinion mining where the discourse of text is short as in forum posts (Ding and Liu, 2010). Hendrickx et al. experimented their coreference resolution with unstructured text in news paper articles, user comments and blog data targeting opinion mining (Hendrickx and Hoste, 2009). Coreference resolution in this domain is restricted to resolve coreferential relations between entities within a discourse of a post. We follow a supervised approach with mention-pair model, learning to identify two mentions are coreferent or not. Mention pairs are constructed from the annotated mentions from the posts. Along with standard set of proven features, grammatical role features and its proposed variants are found to contribute to increase in accuracy. Grammatical role features (Kong et al., 2010; Ng, 2007; Uryupina, 2006) extracted from the dependency parse are intended to capture the characteristics of the human process of coreference resolution, getting the grammatical role of a mention in the corresponding sentence and thus obtaining the relation between the mentions in the pair. Semantic compatibility is a crucial feature in coreference resolution, exploiting named entity (NE) class of mentions. To satisfy the requirements of our domain, NE classes are extended to raga, music concept, music instrument, song.

We have analyzed the importance of dependency parse based grammatical role features, its variants and other features with the limited annotated music forum data available. A rule based chunking implementation is deployed for mention detection. To deal with data insufficiency we have also tried the performance of Bayesian network against SVM in the mention pair classification. This is evaluated with a defined network structure designed to capture some basic known dependencies between features. In this paper we employ a simple network structure with the intention to improve, based on the observations. In our experiments, we observe that Bayesian network has better performance compared to SVM with most of the evaluation metrics.

2 Knowledge Source for Coreference Resolution

Features are computed for a mention pair comprising of potential antecedent mention and anaphoric mention. We make use of a subset of conventional features including the features described in (Soon et al., 2001). String matching (STR_MATCH) and alias (ALIAS) features check for compatibility between the mention with regard to string similarity. These features depend on fuzzy string match-

ing to bypass spelling differences. Same sentence (SAME_SENT) feature checks if both the mentions are in the same sentence and sentence distance gets the number of sentences in between the mentions(SENT_NO). The check for proper noun and pronoun is done for second mention in the pair (PRPN2, PRN2). Features include check for whether a mention is definite (DEF_NP) or demonstrative (DEM_NP).

2.1 Grammatical Role Features

Though the discussed features are significant for showing the coreferent characteristics of a mention pair, the grammatical role of a mention in a discourse and its relation with other mentions are prime features in coreference identification. In a short discourse where the mentions lie in close vicinity, the grammatical role is an important player in deciding coreference when compared to long discourse having coreferent mentions far apart. Apart from analyzing whether a mention in the pair is a subject or object of a sentence, 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 grammatical role of a mention is determined with the help of dependency parse of a sentence obtained from Stanford dependency parser (De Marneffe and Manning, 2008)

These features take into consideration the relevance of a mention with respect to the grammatical role. The coreferent relation between two mentions is dependent on other mentions occurring around the mentions under consideration. So we designed a few other features to capture the behavior of other mentions around, inorder to supplement or weaken the coreferent relation between the mentions in the pair.

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.

Subject mention associated with root verb between (ROOT_SUBJ_BET): This feature is a complement to the previous one, checking for existence of a mention between m_i and m_j having 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.

2.2 Named Entity (NE) Class Feature

Semantic compatibility between the mentions is a critical feature while resolving coreferences (Ng, 2007), making other syntactic features irrelevant on semantic incompatibility. While commonly used NE classes are restricted to person, location, organization etc., in Indian classical music domain it is important to have NE classes like raga, music instrument, music concept, song along with the existing ones.

We follow a dictionary based approach for identification of mention's NE class with the help of entities from Musicbrainz[1]. The mentions are compared against the entities in the dictionary using fuzzy string matching to alleviate the impact of spelling discrepancies. Apart from this, certain heuristics are incorporated (ex. mentions starting with 'Shri' or 'Smt' are person names). Named entity class identification is made offline inorder to support manual curation.

3 Modeling

Since mention-pair model is followed training and testing requires mentions pairs to be formed from the corpus. In a supervised approach training requires positive instances created from mention pairs formed from within a coreferent cluster and negative mention pair instances contain mentions from different clusters. These instances are taken from annotated corpus. While forming mention pairs, the first mention in the pair is chosen to be a non-pronominal mention. An anaphoric mention can never be coreferent with a pronominal mention considering the nature of this corpus. Since the number of negative mention pair far exceeds the number of positive mention pair instances, negative instances are randomly selected from a forum

[1]https://musicbrainz.org/

Feature	Description
First mention subject (SUBJ1)	True, when m_i is a subject of any verb in the sentence
Second mention subject (SUBJ2)	True, when m_j is a subject of any verb in the sentence
First mention object (OBJ1)	True, when m_i is an object of any verb in the sentence
Second mention object (OBJ2)	True, when m_j is an object of any verb in the sentence

Table 1: Basic grammatical role features

post to cap the margin between positive and negative instances.

Test instances are formed from the test file having automatically detected mentions. The accuracy of the system is also dependent on the accuracy of mention detection.

4 Experiment Setup

4.1 Database

Forum	#Posts	#Sent.	#M	#P	#N
Raga & Alapana	143	893	2091	642	1829
Vidwans & Vidushis	180	1219	2749	1247	2742

Table 2: Details of annotated posts. (#Posts= No. of posts #Sent= No. of sentences in the forum. #M= No. of annotated mentions #P= positive mention pairs formed #N= negative mention pairs formed)

The corpus contains coreference annotated forum posts from 2 forums in *rasikas.org*. *Raga & Alapana* has discussions about Carnatic ragas and related concepts and *Vidwans & Vidushis* discusses about Carnatic artistes. Each thread has a title and the posts in the thread discuss the title of the thread. Table 2 shows statistics of annotated forum posts. The annotated data is made available in CoNLL format. Test CoNLL files for validation are also created from the same content by automating mention detection.

4.2 Mention Detection

Mention detection identifies entity boundaries. 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. We depend on Stanford POS tagger for getting POS tags of the corpus(Toutanova et al., 2003). But the POS tagging produced is inaccurate due to noisy text which demands post processing to extract more relevant mentions. Certain proper nouns which are Indian names or Indian classical music terms categorized as nouns by the POS tagger are identified through a dictionary check. Possessive endings marked with different tags are also identified in this step.

Identification of accurate boundaries is challenging due to noisy text with grammatical issues. Making use of knowledge base from web can help in better identification of mention boundaries.

4.3 Evaluation

As explained before training instances are generated from annotated corpus and testing instances from corpus having mentions detected automatically. Experiments are carried out with SVM linear classifier and Bayesian network with predefined network structure. In these domains where the annotated data is scarce and the text is noisy, a Bayesian network with defined structure can work better(Antal et al., 2004). The network structure can incorporate the knowledge available along with the statistical information. Here the Bayesian network will integrate the benefits of both rule based and statistical approaches. A basic network structure is made use as described in fig 1.

We conducted 5-fold cross validation. As the mentions identified through automated mention detection are different from the annotated mentions, the train and test CoNLL content are different in terms of mention boundaries. Still during cross validation the posts considered for training are not included in the testing fold. During 5-fold validation the test mention pairs are classified as coreferent/not coreferent, which are then clustered to form the resultant CoNLL output. We applied best-first clustering(Ng, 2005), where the mention with highest likelihood value is selected as antecedent for an anaphoric mention.

Ablation testing is employed to find weakly per-

Experiments	MUC			B^3			CEAF-M		
	P	R	F	P	R	F	P	R	F
A	33.61	37.44	35.37	42.72	50.82	46.36	36.65	52.58	43.18
B	35.77	54.78	43.19	39.14	58.78	46.98	41.86	60.16	49.35
C	38.16	52.9	44.0	40.02	58.38	47.44	40.84	58.73	48.16

Table 3: Results (P:precision R:recall F:F-measure)
Experiments A: SVM without grammatical role features B: SVM with all features C: Bayes network with all features.

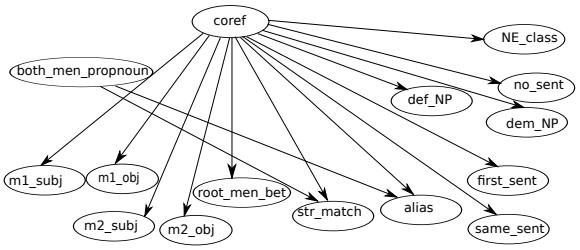

Figure 1: Bayesian network structure depicting dependencies between features

forming features and the most weakly performing 3 features are removed.

5 Results and Discussion

Results are reported in coreference evaluation metrics MUC, B^3 and CEAF-M. Experiment A is without grammatical role features and exp. B clearly indicates the improvement with the grammatical features. Experiment B and C uses all the selected features, using classifiers SVM and Bayes net respectively. As mentioned in section 4.3 the weakly performing features are removed using ablation testing and the results using these features are shown in table 3. The problems with mention detection is one major cause for low accuracy. Even among identified mentions, the mismatch in boundaries is a concern. Analysis of the errors bring forth the major shortcomings and advantages of evaluated classification methods. The problem of semantic incompatible mentions are coreferent with SVM as classifier is almost absent with Bayesian network. Though it contributes well to precision, the recall is seen low compared to SVM because of the relative low importance given to the string matching and alias features.

There are common problems observed with both the classifiers. Despite the hypothesis we had about MEN1_ROOT_SUBJ feature, it is observed that the introduction of this feature reduces accuracy. There are instances of deictic phrases, where the phrase refers to an entity outside the scope of mentions defined in the discourse(Pinkal, 1986). Isolation of deictic phrases can alleviate many false alarms. Certain misclassification occurs at the clustering phase, where the wrong antecedent get selected instead of the correct one even when mention pair with the correct mention is classified as coreferent. Some mentions which are supposed to be singleton are clustered with other clusters because of their linkage with one of the mentions in the cluster.

6 Conclusion and Future Work

This paper focuses on coreference resolution in short discourse of text in Indian classical music. The evaluated mention pair features are expected to capture the specificities of coreferent mentions in short discourses. The devised methods are expected to work well with similar nature forum texts.

Lack of annotated data poses serious problem to classification inspite of the prominent features. Bayesian network exhibits significant improvement in precision despite the small reduction in recall. Bayesian network assures the dominance required for the NE class feature, even though it leads to a few false alarms. The present network structure encodes limited dependencies. A more accurate network structure is evolving based on observations.

Given the fact that semantic/NE class feature has high precedence, accurate extraction of NE class is vital. Even though gender is an important feature, it is not computed due to lack of knowledge sources and methods for computing gender for Indian names. Considering the details of information Freebase posses about each entity, Freebase can aid both these subtasks. Coreference clustering can be further improved incorporating methods to compare belongingness of a mention

to different cluster based on likelihood values between the mention and all the mentions in a cluster, instead of a single mention in the cluster.

References

Peter Antal, Geert Fannes, Dirk Timmerman, Yves Moreau, and Bart De Moor. 2004. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281.

Chinatsu Aone and Scott William Bennett. 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.

Colin Cherry and Shane Bergsma. 2005. An expectation maximization approach to pronoun resolution. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 88–95. Association for Computational Linguistics.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

Xiaowen Ding and Bing Liu. 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.

Iris Hendrickx and Veronique Hoste. 2009. Coreference resolution on blogs and commented news. In *Anaphora Processing and Applications*, pages 43–53. Springer.

Fang Kong, Guodong Zhou, Longhua Qian, and Qiaoming Zhu. 2010. Dependency-driven anaphoricity determination for coreference resolution. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 599–607. Association for Computational Linguistics.

Joseph F McCarthy and Wendy G Lehnert. 1995. Using decision trees for coreference resolution. *arXiv preprint cmp-lg/9505043*.

Vincent Ng. 2005. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 157–164. Association for Computational Linguistics.

Vincent Ng. 2007. Semantic class induction and coreference resolution. In *ACL*, pages 536–543.

Manfred Pinkal. 1986. Definite noun phrases and the semantics of discourse. In *Proceedings of the 11th coference on Computational linguistics*, pages 368–373. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 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.

rasikas. 2015. Rasikas.org. `http://www.rasikas.org`.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Mohamed Sordo, Joan Serrà Julià, Gopala Krishna Koduri, and Xavier Serra. 2012. Extracting semantic information from an online carnatic music forum. 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).

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Olga Uryupina. 2006. Coreference resolution with and without linguistic knowledge. In *Proceedings of LREC*, pages 893–898.