Improved Best-First Clustering for Coreference Resolution in Indian Classical Music Forums

Joe Cheri Ross and Pushpak Bhattacharyya

Dept. of Computer Science & Engg:, Indian Institute of Technology Bombay, Mumbai
{joe, pb}@cse.iitb.ac.in

Abstract. Clustering step in the mention-pair paradigm for coreference resolution, forms the chain of coreferent mentions from the mention pairs classified as coreferent. Clustering methods including best-first clustering considers each antecedent candidate individually, while selecting the antecedent for an anaphoric mention. Here we introduce an easy-to-implement modification to best-first clustering to improve coreference resolution on Indian classical music forums. This method considers the relation between the candidate antecedents along with the relation between the anaphoric mention and the candidate antecedent. We observe a modest but statistically significant improvement over the best-first clustering for this dataset.

Keywords: coreference resolution, information extraction, Indian classical music

1 Introduction

Coreference resolution is the task of finding mentions in a discourse referring to the same entity and grouping them into a set [1]. The motivation behind improving coreference resolution on Indian classical music forums is to be improve relation extraction from these forums, thus contributing to meta information in knowledge base for Indian classical music. Many of the forums and blogs on Indian classical music are rich source of information. Rasikas.org [2] forum considered for this study, has discussions in English on different topics in Carnatic music (sub-genre of Indian classical music). Considering the relevance of extractable information from this forum to the knowledge base for Indian classical music, coreference resolution is vital in improving extraction of relations.

The coreference resolution approach described in this paper is based on mention-pair model [3, 4], where the classification of mention pairs is followed by clustering to form chain of coreferent mentions. The classification approach is hybrid with a rule-based sieve and machine learning based classifier. Pair wise classification decisions are utilized for partitioning coreferent mentions in clustering [5]. There are a few existing approaches for clustering. To find the antecedent of an anaphoric mention, best-first clustering considers all the mention pairs classified as coreferent with the anaphoric mention. The best mention

pair is picked to find the right antecedent, based on the classification confidence associated with the mention pair [6, 4]. The closest-first approach selects the closest preceding coreferent mention in the discourse as the antecedent [7]. Aggressive-merge approach selects all coreferent mentions to the anaphoric mention and make it part of the same coreferent chain [3]. Our method introduces an improvement over best-first clustering.

In the mention-pair model, mention pairs are formed between an anaphoric mention (m_{ana}) and candidate antecedent mentions which precede the anaphoric mention in the discourse. Mention pair classification classifies these mention pairs as coreferent or not. From the coreferent mention pairs involving the anaphoric mention, best-first clustering selects the antecedent (m_{ant}) from the mention pair having the highest classification confidence score associated with it. The probability estimate of mention-pair classification serves for the confidence score.

$$m_{ant} = \underset{m_c \in candidate\ antecedents}{argmax} P((m_c, m_{ana})) \quad (1)$$

Where $P((m_c, m_{ana}))$ denotes the classification probability estimate associated with the mention pair (m_c, m_{ana}). The modification to best-first clustering proposed in this paper, 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.

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 another mention which has no coreferent relation with other candidates. 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 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, bold line→strong coreference relation). The strong coreference relation between the candidates *Snehapriya* and *raga snehapriya* makes them better candidates over others. 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 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 attached 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.

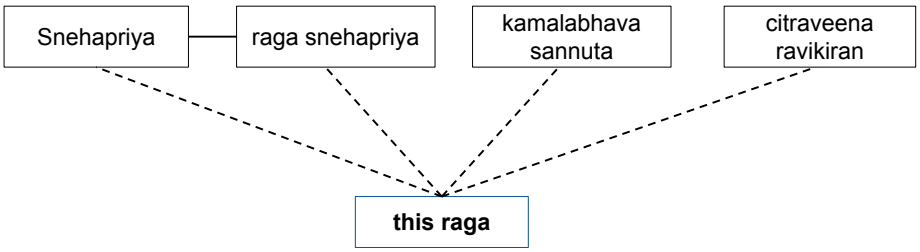

Fig. 1. An example scenario of antecedent selection taken from a forum post

This clustering method identifies all such support mentions for a candidate antecedent and computes the **support score** (refer Algorithm 1). 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 2). This confidence score replaces the probability estimate in Equation 1 to find the best antecedent for an anaphoric mention.

2.1 Algorithm

Algorithm 1 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 1 describes the method to compute the support score for a candidate antecedent given an anaphoric mention (m_{ana}). The support score ($supp$) is computed for all candidate antecedents of this anaphoric mention. The method takes the mention pair involving a candidate antecedent (ex. (*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 (ex. *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 1 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

$$confidence\ score = \lambda P_e + (1 - \lambda) supp, \lambda \in (0, 1) \tag{2}$$

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.

2.2 Dynamic λ

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 3 takes the probability estimate value associated with the mention pair classification as its classification confidence.

$$\lambda = kP_e, k \in (0, 1) \tag{3}$$

where k is a constant. An alternate method is devised to decide classification confidence. 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

$$\lambda = \frac{1}{1 + clsf\_var} \tag{4}$$

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.

3 Dataset: Rasikas.org

The coreference annotated dataset contains forum posts from Rasikas.org. This is a prominent discussion forum for Carnatic music, which is the classical music of south India. The main topics of discussion in the forum includes raga [8], *tala* (rhythm), *vidwans & vidushis* (musicians), *vaggeyakaras* (composers), *kutcheri* (concert) reviews & recordings, album reviews, etc. Table 1 shows the details of this dataset. This forum is a rich source of information and listeners' opinions in the mentioned topics.

Forum	**#Posts**	**#Sent.**	**#Mentions**
Raga & Alapana	300	2093	3630
Vidwans & Vidushis	587	3045	10884
Vaggeyakaras	325	2339	4421

Table 1. Details of annotated posts.

Each forum post is a short discourse text comprising 4-5 sentences on an average. 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.

4 Experiments & Results

Experiments		**MUC**			**B^3**			**CEAFe**			**CoNLL**
		P	R	F	P	R	F	P	R	F	
Neural Net	BF	55.45	62.35	58.38	54.84	65.36	59.44	50.62	60.75	54.88	57.56
	supp-BF	55.67	62.81	58.70	54.92	65.91	59.70	50.75	60.76	54.96	**57.79**
	supp-BF-1	55.78	62.71	58.72	55.00	65.86	59.74	50.74	60.90	55.02	57.83
	supp-BF-2	55.54	62.71	58.57	54.89	65.71	59.61	50.71	60.71	54.93	57.70
SVM (RBF)	BF	48.42	64.96	55.28	49.66	66.02	56.56	54.83	57.09	55.45	55.76
	supp-BF	48.93	65.57	55.84	49.77	67.01	57.00	55.01	57.29	55.64	**56.16**
	supp-BF-1	48.92	65.56	55.83	49.76	67.00	57.00	55.01	57.29	55.64	56.16
	supp-BF-2	48.77	65.35	55.65	49.73	66.73	56.88	54.99	57.25	55.61	56.05

Table 2. Results with different classifiers (P,R,F)→ (Precision, R:Recall, F:F-measure), CoNLL:CoNLL Score. CoNLL score of significant improvements are in bold.

Our system follows the mention-pair model with a machine learning approach. Conventional features and the features which are found to be more important for this domain are employed [9]. We employ k-fold (5 folds) cross-validation to make the maximum utilization of available annotated dataset. The consistency of the methods is validated across 2 different classifiers, *viz.*, Multi-layered Feed-Forward Neural Network (Neural Net) and Support Vector Machine (SVM). Effectively, validation of the system is done with predicted mentions. Results are reported with MUC [10], B^3 [11] and CEAFe [12] metrics. The average of F-measures from all these metrics is taken as CoNLL Score.

Table 2 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 $conf\_thresh$=0.9 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, experiment supp-BF produces a noticeable improvement in accuracy compared to best-first clustering. Figure 2 shows the reduction in recall errors for nominal and pronoun anaphora types in supp-BF compared to BF. 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 improvement in accuracy of supp-BF-1 over supp-BF is very small. supp-BF-2 produces no improvement in accuracy compared to supp-BF and supp-BF-1, but better compared to the baseline best-first.

The significance of the accuracy improvement is tested with a paired-t test on CoNLL scores [14]. For this, the test set is divided into 20 sub-samples

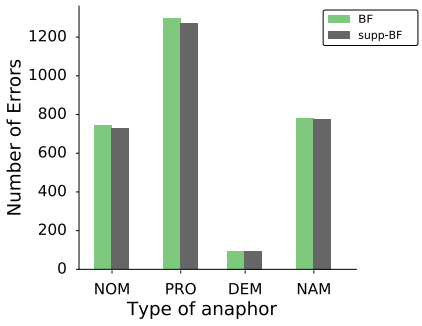

Fig. 2. Anaphora type wise comparison of errors between BF and supp-BF (Produced by Cort error analyzer [13])

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.

5 Conclusion and Future Work

This paper 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 baseline but not significant over the basic variant.

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.

References

1. Jie Cai and Michael Strube. End-to-end coreference resolution via hypergraph partitioning. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 143–151. Association for Computational Linguistics, 2010.
2. rasikas. Rasikas.org, 2005.
3. Joseph F McCarthy and Wendy G Lehnert. Using decision trees for coreference resolution. *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995.

4. Chinatsu Aone and Scott William Bennett. 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, 1995.
5. Vincent Ng. 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, 2010.
6. Vincent Ng and Claire Cardie. 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, 2002.
7. Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001.
8. Saraswathy Bhagyalekshmy. *Ragas in Carnatic music*. South Asia Books, 1990.
9. Joe Cheri Ross and Pushpak Bhattacharyya. Coreference resolution to support ie from indian classical music forums. *Recent Advances in Natural Language Processing*, page 91, 2015.
10. Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics, 1995.
11. Amit Bagga and Breck Baldwin. 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, 1998.
12. Xiaoqiang Luo. 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, 2005.
13. Sebastian Martschat, Thierry Göckel, and Michael Strube. 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, 2015.
14. Jie Cai and Michael Strube. 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, 2010.