Pronominal Anaphora Resolution in Konkani language incorporating Gender Agreement

Poonam A. Navelker[1,2]
[1]Goa Business School, Goa University
Taleigao 403207, India
[2]SES's Sridora Caculo College of Commerce and Management Studies
Mapusa 403507, India
dcst.poonam@unigoa.ac.in

Jyoti D. Pawar[1]
[1]Goa Business School, Goa University
Taleigao 403207, India
jdp@unigoa.ac.in

Abstract

Konkani is a low-resource language, spoken mainly on the central west coast of India. Approximately 2.3 million people speak Konkani (Office of the Registrar General Census Commissioner, India, 2011). It is also the official language of the state of Goa. It belongs to the Southern Indo-Aryan language group. The official Script for writing the Konkani language is Devanagari. Despite this, being a low-resource language has hampered its development on the digital platform, Konkani has yet to significantly impact its digital presence. To improve this situation, contribution to Natural Language Understanding in the Konkani language is important. This paper aims to resolve pronominal anaphora in the Konkani language using a rule-based method incorporating gender agreement. This is required in NLP applications like text summarization, machine translation, and question-answering systems. While research on English and other foreign languages, as well as Indian languages like Tamil, Kannada, Malayalam, Bengali, and Marathi, have been done, no work has been done on the Konkani language thus far. This is the very first attempt made to resolve anaphora in Konkani.

1 Introduction

In linguistics, anaphora resolution is used. It resolves the reference pointing to the earlier entity in the discourse. Reference used to point to the entity is called an "Anaphor" and the entity pointed by the reference is called an "Antecedent".

Example 1:

गौतमान नवी गाडी घेतली. तो खूब खोशी आसा.
Here, 'तो' refers to 'गौतम' in the first sentence.
So, 'गौतम' is the antecedent of the anaphor 'तो'.

There are different types of anaphora. Some of the types are Pronominal anaphora, One anaphora, Do-so anaphora, sentential-it anaphora, and gapping. Anaphora resolution varies across different languages. Many approaches have been used for anaphora resolution. This paper focuses on a rule-based method to resolve pronominal anaphora incorporating gender agreement. The paper is organized as follows. Section 2 contains a Literature Review. Section 3 explains the Pronominal Anaphora Resolution in Konkani, the basic cases implemented, and our proposed rule-based algorithm incorporating gender agreement. Section 4 states the Results and Discussion. Section 5 contains the Conclusion and Future work.

2 Literature Review

Anaphora resolution has evolved in the past few decades. Many approaches have been used in resolving anaphora like, rule-based, machine learning-based, knowledge-based, and so on. The salience feature-based approach was used by (Lappin and Leass, 1994) for the English and German languages. Simple noun-based rules and part of speech tagger were used by (Mitkov, 1998). For Indian languages, work has been carried out for languages like, Hindi, Bengali, Malayalam, and Tamil. Researchers have started working recently on the Marathi language. For the Malayalam and Hindi languages, a rule-based approach VASISTH was used by (sob, 2002). A hybrid approach for anaphora resolution in Hindi using a dependency parser and decision tree classifier was used by (Dakwale et al., 2013). In 2014, (Lalitha Devi et al., 2014) came up with a Generic anaphora resolution engine for Indian languages . It used a machine learning approach. In 2015,(Jonnalagadda and Mamidi, 2015)came up with a rule-based pronominal anaphora resolution for Telugu dialogues. In 2019, (Khandale and Mahender, 2019) came up with a rule-based anaphora resolution for Marathi sentences. In 2022, (Khandale and Mahender, 2022) came up

with a rule-based anaphora resolution approach in Marathi focusing on binding theory.

3 Pronominal Anaphora Resolution in Konakni

There are different types of pronouns in Konkani as follows:

1.	Personal	हांव, आमी, तुमी (Haav, Aami, Tumi)
2.	Possessive	म्हजे, म्हजी, तुजे, तुजी (Mhaje, Mhaji, Tuje, Tuji)
3.	Demonstrative	तो, ती, तें (To, Ti, Te)
4.	Reflexive	आमी, तुमी, आमकां, तुमकां, स्वता (Aami, Tumi, Aamka, Tumka, Swata)
5.	Reciprocal	एकमेकांक, एकमेकांचो, आपल्याक (Ekmekank, Ekmekancho, Aaplyak)
6.	Interrogative	किदें, कशें, कोण, किद्याक (Kide, Kashe, Kon, Kidyak)

In this paper, we are focusing on the demonstrative pronouns and their variations as follows:

तो, ती, तें, ताणें, तिणें, तांणी, ताचें, तिचे

3.1 Our Approach

Our approach to anaphora resolution is as follows:
1. Create the dataset
2. POS tag data
3.Apply a rule-based algorithm to identify the antecedent of the anaphor

3.2 Dataset

Konkani is a low-resource language. Researchers faced problems in getting a corpus containing pronominal anaphora. Therefore, a set of 146 sentences was generated based on day-to-day conversations. It contains demonstrative pronouns and their variations. A pair of 71 sentences contained anaphor and antecedents. The remaining were single sentences having anaphors and antecedents. The dataset contains demonstrative pronoun and their variations as follows:

तो, ती, तें, ताणें, तिणें, तांणी, ताचें, तिचे

3.3 Part of Speech Tagger

The sentences from the dataset created were POS tagged using the Konkani Online POS tagger created by Annie Rajan and Arshad Shaikh which used deep learning models (Rajan and Shaikh, n.d.). This POS tagger did not generate accurate results, this is the only POS tagger available for the Konkani language. We manually corrected the wrongly tagged POS words. The corrected POS-tagged words were stored in a different file.

3.4 Methodology: Rule-Based Algorithm

For a resource-poor language, rule-based algorithms are best to start with. So we have used a rule-based pronoun resolution algorithm. Rules were formed based on the demonstrative pronoun usage in the Konkani language, incorporating gender agreement. Two sentences having the following variations are used:

Case 1:

Sentence 1:containing one common noun or one proper noun

Sentence 2: containing one pronoun

Rule 1:

In such a case, the pronoun in the second sentence points to either a common noun or a proper noun. This is the simplest case.

Example 2:

Sentences	अमित हुशार आसा. शाळेंत तो पयलो येता.
POS Tagged data	((([('अमित', 'N_NNP'), ('हुशार', 'JJ'), ('आसा', 'V_VM_VF')])) ((([('शाळेंत', 'N_NN'), ('तो', 'PR_PRP'), ('पयलो', 'QT_QTO'), ('येता', 'V_VM_VF')]))
Output	The antecedent of the pronoun 'तो' is the noun 'अमित'.

Figure 1: Anaphora Resolution of sentence 1 containing one proper noun

Case 2:

Sentence 1: containing one proper noun and one or more common nouns

Sentence 2: containing one pronoun

Rule 2: In such a case, the pronoun in the second sentence points to the proper noun in the first sentence.

Example 3:

Sentences	राजून बाजारांतल्यान जिलबी घेतली. ताणें घरा हाडून आवयक दिली.
POS Tagged data	(([('राजून', 'N_NNP'), (' बाजारांतल्यान ', 'N_NN'), ('जिलबी', 'N_NN'), ('घेतली', 'V_VM_VF')])) (([('ताणें', 'PR_PRP'), ('घरा', 'N_NN'), ('हाडून', 'V_VM_VNF'), ('आवयक', 'N_NN'), ('दिली', 'V_VAUX_VF')]))
Output	The antecedent of the pronoun 'ताणें' is the noun 'राजून'.

Figure 2: Anaphora Resolution of sentence 1 containing one proper noun and one or more common nouns

Case 3:
Sentence 1: containing one proper noun and one common noun
Sentence 2: containing one pronoun
Applying Rule 2 could only solve certain types of cases. It failed to identify cases like
Example 4:

Sentences	बाबून लाडू खालो. तो गोड आशिल्लो.
POS Tagged data	(([(' बाबून ', 'N_NNP'), ('लाडू', 'N_NN'), ('खालो', 'V_VM_VF')])) (([('तो', 'PR_PRP'), ('गोड', 'JJ'), ('आशिल्लो', 'V_VM_VF')]))
Output	The antecedent of the pronoun 'तो' is the noun' बाबून '.

Figure 3: Anaphora Resolution when the pronoun is followed by an Adjective in Sentence 2

To resolve the above case correctly, Rule 3 can be applied.
Rule 3:
If
(Proper noun is followed by a common noun in the first sentence and pronoun is followed by an adjective in the second sentence)
Then
(Pronoun refers to the common noun)

This could resolve Example 4 correctly.After applying Rule 3, The output is:

The antecedent of the pronoun 'तो' is the noun 'लाडू'.

Case 1, Case 2, and Case 3 are very basic cases and have been addressed in the Marathi language for a single sentence (Khandale and Mahender, 2019)

3.5 Our Contribution

Konkani is similar in some aspects to the Marathi language, so Rule 1, Rule 2, and Rule 3 could be adopted. These basic resolutions are needed as this is the first work being carried out for the Konkani language. Once these basic anaphora rules are applied, we can solve more complex cases. Also, the gaps mentioned in the above paper are addressed here.
We could identify exceptions to Case 3 that could not resolve the correct antecedent using Rule 3.
Example 5:

Sentences	निधीन सर्कस पळयली. ते अजाप जाले.
POS Tagged data	(([('निधीन', 'N_NNP'), ('सर्कस', 'N_NN'), ('पळयली', 'V_VM_VF')])) (([('ते', 'PR_PRP'), ('अजाप', 'JJ'), ('जाले', 'V_VM_VF')]))
Output	The antecedent of the pronoun 'ते' is the noun ' सर्कस '.

Figure 4: Exception to Rule 3

The above-generated output is not correct. It should be

The antecedent of the pronoun 'ते' is the noun ' निधीन '.

Rule 3 needs to be changed to get the correct output. To resolve the above case, rule 3 is modified to incorporate gender agreement.
Case 3:
Sentence 1: containing one proper noun and one common noun
Sentence 2: containing one pronoun

Rule 4: Incorporating Gender Agreement

Explanation: As per the given algorithm in Figure 5, we first check the structure of the sentences. If in Sentence 1, the proper noun is followed by the common noun (N_NN), and in Sentence 2, the pronoun (PR_PRP)is followed by an adjective (JJ), then we check if it satisfies the following condition. The condition is, that if the ending characters of verbs in Sentences 1 and 2 end with the pairs mentioned in the algorithm, it means there is gender agreement of verbs. In such case, the antecedent is the common noun (N_NN) from

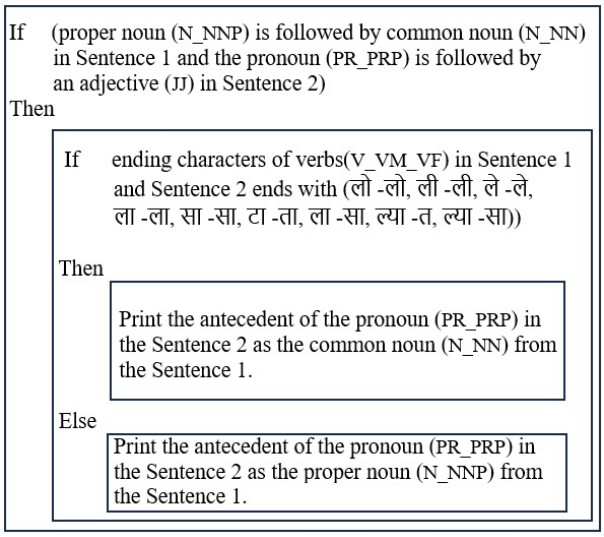

Figure 5: Our proposed algorithm to resolve anaphora incorporating gender agreement

the Sentence 1. On the other hand, if there is no gender agreement of verbs, then the antecedent is the proper noun (N_NNP) from the Sentence 1.

Figure 6 and Figure 7 show the output generated by applying our rule-based anaphora resolution algorithm incorporating gender agreement.

Example 6:

Sentences	निधीन सर्कस पळयली. ते अजाप जाले.
POS Tagged data	(([('निधीन', 'N_NNP'), ('सर्कस', 'N_NN'), ('पळयली', 'V_VM_VF')])) (([('ते', 'PR_PRP'), ('अजाप', 'JJ'), ('जाले', 'V_VM_VF')]))
Output	The antecedent of the pronoun 'ते' is the noun ' निधीन '.

Figure 6: Anaphora Resolution when gender agreement rule does not match

Here, the gender agreement of verbs in both sentences does not match. Since gender agreements do not match, the antecedent is the proper noun from the first sentence.

Example 7:

Here, the gender agreement of verbs in both sentences matches. Since gender agreements match, the antecedent is the common noun from the first sentence.

Sentences	ताईबायन मुदी घाल्ली. ती चकचकीत आशिल्ली.
POS Tagged data	(([('ताईबायन', 'N_NNP'), ('मुदी', 'N_NN'), ('घाल्ली', 'V_VM_VF')])) (([('ती', 'PR_PRP'), ('चकचकीत', 'JJ'), ('आशिल्ली', 'V_VM_VF')]))
Output	The antecedent of the pronoun 'ती' is the noun 'मुदी'.

Figure 7: Anaphora Resolution when gender agreement rule matches

Case 4:
Sentence 1: containing two proper nouns and one pronoun.

Rule 5:

In such a case, check the ending character of each proper noun. If a proper noun ends with 'न' (excluding those ending with 'क'), then the pronoun refers to the proper noun ending with 'न'.

Example 8:

Sentences	रश्मिन निताक बाजारान व्हेले आनि ती खूब थकली.
POS Tagged data	(('रश्मिन', 'N_NNP'), ('निताक', 'N_NNP'), ('बाजारान', 'QT_QTC'), ('व्हेले', 'V_VM_VNF'), ('आनि', 'QT_QTF'), ('ती', 'PR_PRP'), ('खूब', 'N_NN'), ('थकली', 'V_VM_VF'))
Output	The antecedent of the pronoun 'ती' is the noun रश्मि.

Figure 8: Anaphora Resolution when sentence contains two proper nouns

Here we have considered the simplest form of a sentence containing 2 proper nouns.

4 Results and Discussion

The evaluation measure used to test the system is accuracy. Accuracy is used to measure the performance of the system in Natural language Processing.

$$\text{Accuracy} = \frac{\text{No. of correct predictions}}{\text{No. of correct predictions+No. of incorrect predictions}} \times 100$$

The rules were implemented using Python programming. The system performed well giving an accuracy of 89.33%. The sentences whose

anaphora was not resolved correctly had a structure of Sentence 1 having 2 common nouns and Sentence 2 having one pronoun. We have not addressed this sentence structure in this paper. As long as the given structure and rules mentioned in this paper were used, the anaphora of the sentences was resolved correctly.

5 Conclusion and Future work

Through this work, we have created a rule-based anaphora resolution algorithm incorporating gender agreement for Konkani language. Anaphora resolution focussed on the demonstrative pronouns and their variations. Four cases with different sentence structures were addressed. We also encountered sentence structures having 2 or more common nouns and no proper noun whose anaphora could not be resolved. In the future, we want to design a rule which will address the limitation and also build up our dataset, so that we can test the system in a better way. We also want to address more complex cases of type Case 4.

References

2002. *VASISTH: An Anaphora Resolution System for Hindi and Malayalam*. Indian Society for Technical Education, Indian Institute of Technology Kanpur, India.

Praveen Dakwale, Vandan Mujadia, and Dipti M Sharma. 2013. A hybrid approach for anaphora resolution in Hindi. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 977–981, Nagoya, Japan. Asian Federation of Natural Language Processing.

Hemanth Reddy Jonnalagadda and Radhika Mamidi. 2015. Resolution of pronominal anaphora for telugu dialogue. In *Proceedings of the 12th International Conference on Natural Language Processing (ICON 2015)*.

Kalpana Khandale and C. Namrata Mahender. 2019. Rule-based design for anaphora resolution of marathi sentence. In *Proceedings of the 5th International Conference for Convergence in Technology (I2CT 2019)*, Pune, India. IEEE.

Kalpana B. Khandale and C. Namrata Mahender. 2022. Resolving anaphora in marathi language. *International Journal of Computer Science and Information Technology*, 17(8).

Sobha Lalitha Devi, Vijay Sundar Ram, and Pattabhi RK Rao. 2014. A generic anaphora resolution engine for Indian languages. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1824–1833, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Office of the Registrar General Census Commissioner, India. 2011. Census of india 2011. Accessed: 2024-09-01.

Annie Rajan and Arshad Shaikh. n.d. PoS Tagger for Konkani. Tool developed by Annie Rajan and Arshad Shaikh. Licensed under CC BY-NC 4.0. Dataset is from ILCI. Contact: ann_raj_2000@yahoo.com.