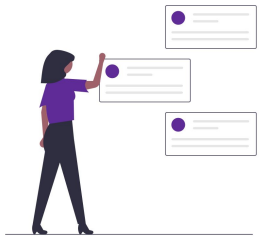
Pronominal Anaphoric Resolution in Malayalam

Enfa Rose George , UG Scholar
Under the guidance of Mr.Mathews Abraham, Assistant Professor
Rajagiri School Of Engineering and Technology

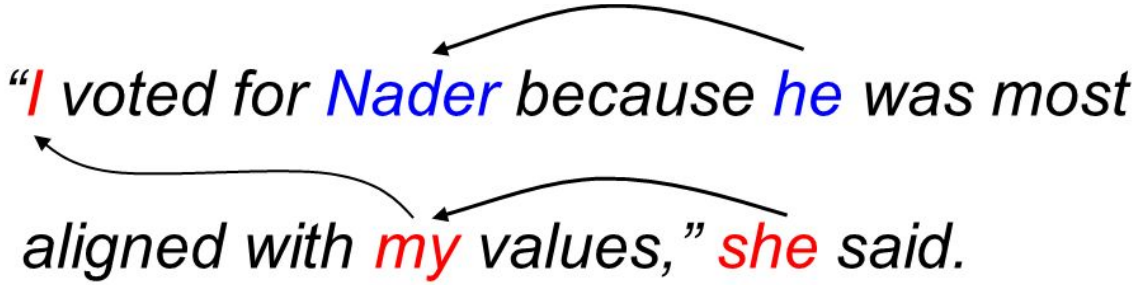
Natural Language Processing

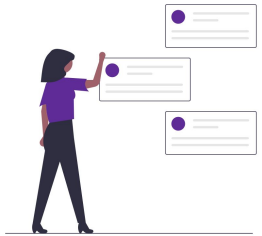
- Natural Language Processing is broadly defined as the automatic manipulation of natural language, like speech and text, by software.
- Some of the major tasks in NLP are
  - Speech recognition
  - Natural language understanding
  - Natural language generation



Coreference Resolution

- Coreference resolution is the task of ***finding all expressions*** that ***refer*** to the ***same entity*** in a text.





Types Of Coreference

- **Anaphora**

The music$_i$ was so loud that **it**$_i$ couldn't be enjoyed. –The anaphor *it* follows the expression to which it refers (its antecedent).

- **Cataphora**

If **they**$_i$ are angry about the music, **the neighbors**$_i$ will call the cops. – The cataphor *they* precedes the expression to which it refers (its postcedent).

- **Split antecedents**

Carol$_i$ told **Bob**$_j$ to attend the party. **They**$_{i+j}$ arrived together. – The anaphor *they* has a split antecedent, referring to both *Carol* and *Bob*.

- **Coreferring noun phrases**

The project leader$_i$ is refusing to help. **The jerk**$_i$ thinks only of himself. – Coreferring noun phrases, whereby the second noun phrase is a predication over the first.

Scope & Application

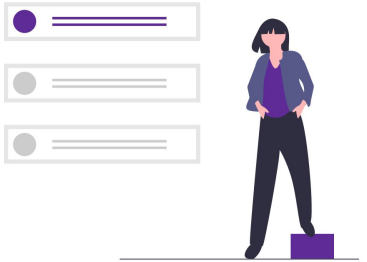
- CR is the ultimate test for human like AI. CR systems is the basis for Winograd Schema Challenge, the test for AI systems that beat the Turing test.
- This is because it requires ***Knowledge and commonsense reasoning*** to solve them.
- The choices of "feared" and "advocated" turn the schema into its two instances:

  The city councilmen refused the demonstrators a permit

  because they **feared** violence.

  The city councilmen refused the demonstrators a permit

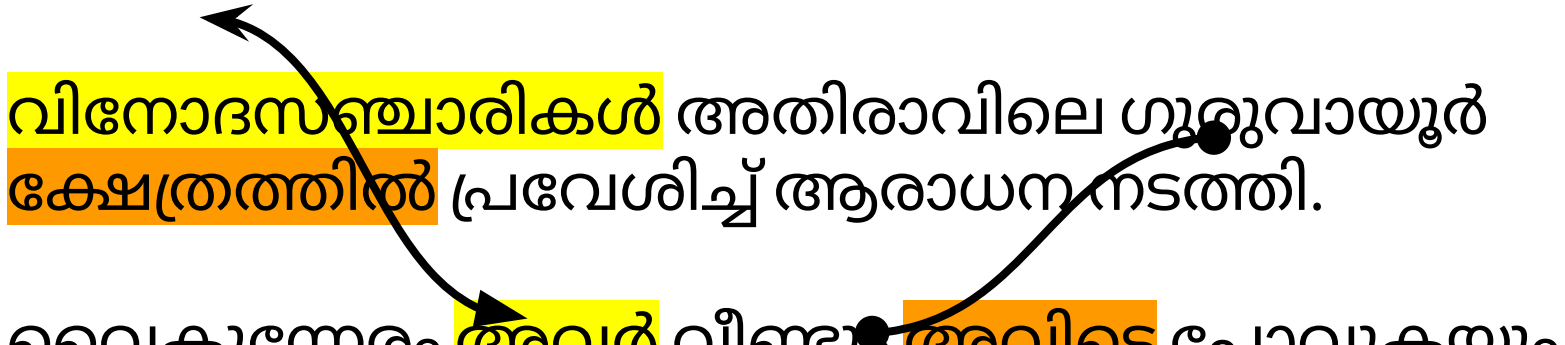
  because they **advocate**d violence.



Scope & Application

- Critical for improvement of tasks like
    - Abstract Document summarization
    - Question answering
    - Information extraction.

The Objective Of the Paper



വിനോദസഞ്ചാരികൾ അതിരാവിലെ ഗുരുവായൂർ ക്ഷേത്രത്തിൽ പ്രവേശിച്ച് ആരാധന നടത്തി.

വൈകുന്നേരം അവർ വീണ്ടും അവിടെ പോവുകയും ദൈവത്തെ ആരാധിക്കുകയും ചെയ്തു

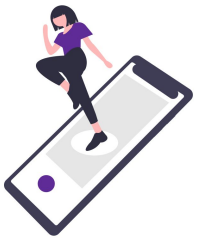
The Method

How do we approach the problem?

General Idea

- Algorithms intended to resolve references commonly ***look first for the nearest preceding individual that is compatible with the referring expression (Hobbs Algorithm)***

- The literature in Malayalam on the topic uses **inhouse data sets to built parsers and taggers**.
- There are a lot of handwritten rules, and are **noticed to be heavy**.

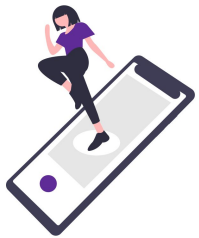


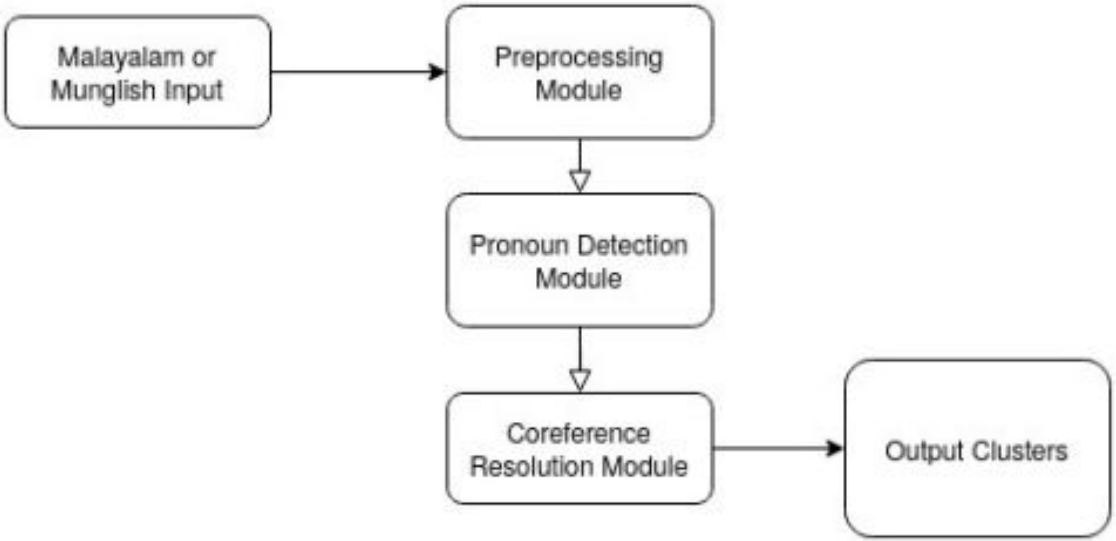
Design and Implementation Constraints

- Malayalam is a **morphologically rich language** which means we have more features to take into consideration

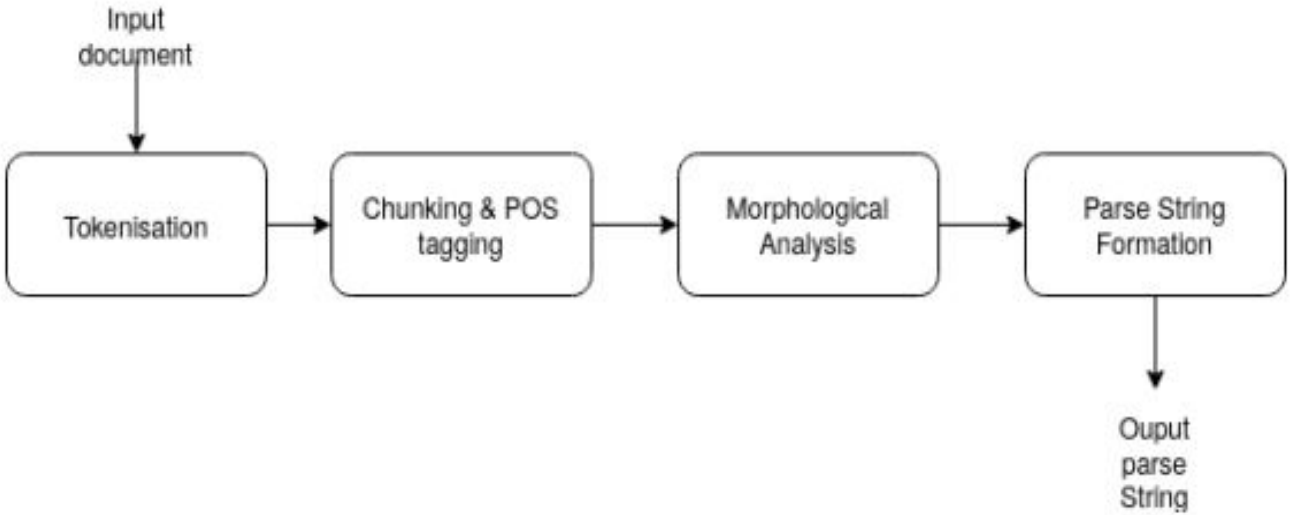
- The performance of Parsers and Taggers publicly available is a direct bottleneck.

- Resource - poor language





Basic Architecture

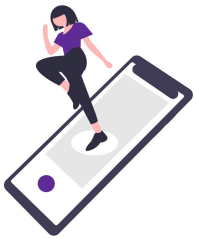


Preprocessing Module

Preprocessing Module

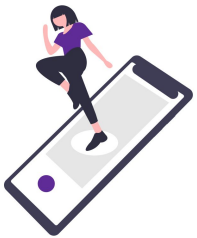
- Tokenization ie ***splitting our document to sentences*** using polyglot library which uses Unicode Text Segmentation Algorithm.

['വിനോദസഞ്ചാരികൾ അതിരാവിലെ ഗുരുവായൂർ ക്ഷേത്രത്തിൽ പ്രവേശിച്ച് ആരാധന നടത്തി .', 'വൈകുന്നേരം അവർ വീണ്ടും അവിടെ പോവുകയും ദൈവത്തെ ആരാധിക്കുകയും ചെയ്തു .']



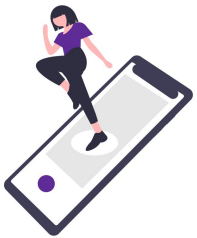
Preprocessing Module

- ***Chunking*** (Converting Sentences into phrases) & minimal ***Parts of Speech tagging*** using Devadath's Shallow Parser.
- ***Morphological Analysis*** done using HFST implementation mlmorph
- We generate a knowledge parse string combining all of this info.



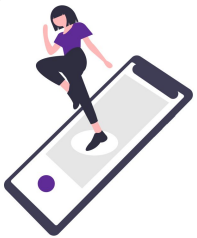
Preprocessing Module

- *[[['വിനോദസഞ്ചാരികൾ ', 'N', 'ADJ', 'N', 'PL', 'N__NN', 'B-NP'], ['അതിരാവിലെ ', 'ADJ', 'NP', 'N__NN', 'B-NP'], ['ഗുരുവായൂർ ', 'NP', 'N__NN', 'B-NP'], ['ക്ഷേത്രത്തിൽ ', 'N', 'LOCATIVE', 'N__NN', 'B-NP'], ['പ്രവേശിച്ച്', 'V__VM__VNF', 'B-VGNF'], ['ആരാധന', 'N', 'N__NN', 'B-NP'], ['നടത്തി', 'V__VM__VF', 'B-VGF'], ['.', 'RD__PUNC', 'B-BLK']], [['വൈകുന്നേരം ', 'NP', 'N__NN', 'B-NP'], ['അവർ', 'PRN', 'PR__PRP', 'B-NP'], ['വീണ്ടും', 'RB', 'B-RBP'], ['അവിടെ', 'CNJ', 'PR__PRP', 'B-NP'], ['പോവുകയും ', 'V__VM__VINF', 'B-VGINF'], ['ദൈവത്തെ', 'N', 'ACCUSATIVE', 'N__NN', 'B-NP'], ['ആരാധിക്കുകയും ', 'V__VM__VINF', 'B-VGINF'], ['ചെയ്തു', 'V__VM__VF', 'B-VGF'], ['.', 'RD__PUNC', 'B-BLK']]]*



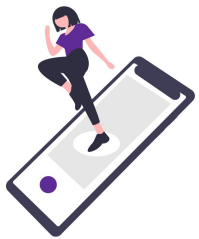
Preprocessing Module- Sieve

- Do some basic candidate invalidation/screening.
- Verb Phrases cannot be a possible candidate.
- *[[['വിനോദസഞ്ചാരികൾ ', 'N', 'ADJ', 'N', 'PL', 'N__NN', 'B-NP'], ['അതിരാവിലെ ', 'ADJ', 'NP', 'N__NN', 'B-NP'], ['ഗുരുവായൂർ ', 'NP', 'N__NN', 'B-NP'], ['ക്ഷേത്രത്തിൽ ', 'N', 'LOCATIVE', 'N__NN', 'B-NP'], ['പ്രവേശിച്ച്', 'V__VM__VNF', 'B-VGNF'], ['ആരാധന', 'N', 'N__NN', 'B-NP'], ['നടത്തി', 'V__VM__VF', 'B-VGF'], ['.', 'RD__PUNC', 'B-BLK']], [['വൈകുന്നേരം ', 'NP', 'N__NN', 'B-NP'], ['അവർ', 'PRN', 'PR__PRP', 'B-NP'], ['വീണ്ടും ', 'RB', 'B-RBP'], ['അവിടെ', 'CNJ', 'PR__PRP', 'B-NP'], ['പോവുകയും ', 'V__VM__VINF', 'B-VGINF'], ['ദൈവത്തെ', 'N', 'ACCUSATIVE', 'N__NN', 'B-NP'], ['ആരാധിക്കുകയും ', 'V__VM__VINF', 'B-VGINF'], ['ചെയ്തു', 'V__VM__VF', 'B-VGF'], ['.', 'RD__PUNC', 'B-BLK']]]*
-



Preprocessing Module - Sieve

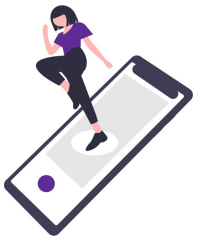
- ['(S (CHUNK (N (NN (N (ADJ (N (PL വിനോദസഞ്ചാരികൾ))))))) (CHUNK (N (NN (ADJ (NP അതിരാവിലെ))))) (CHUNK (N (NN (NP ഗുരുവായൂർ)))) (CHUNK (N (NN (N (LOCATIVE ക്ഷേത്രത്തിൽ))))) (CHUNK (N (NN (N ആരാധന))))) ', '(S (CHUNK (N (NN (NP വൈകുന്നേരം)))) (CHUNK (PR (PRP (PRN അവർ)))) (CHUNK (PR (PRP (CNJ അവിടെ)))) (CHUNK (N (NN (N (ACCUSATIVE ദൈവത്തെ)))))) ']



Pronoun Detection

```
obj.returnPronounText()
```

```
{(1,
  1): 'അവർ in the sentance വൈകുന്നേരം അവർ വീണ്ടും അവിടെ പോവുകയും ദൈവത്തെ ആരാധിക്കുകയും ചെയ്യ .',
 (1,
  3): 'അവിടെ in the sentance വൈകുന്നേരം അവർ വീണ്ടും അവിടെ പോവുകയും ദൈവത്തെ ആരാധിക്കുകയും ചെയ്യ .'}
```

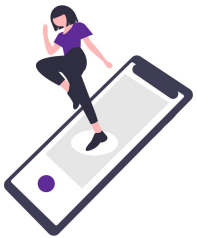


Naive CR Algorithm - slight variation of Hobbs

- **Do *left to right* depth first parsing *sentence wise***, looking for NP - Proper Nouns and then NN - Common Nouns candidates.

```
obj.returnSolutionCandidates()
```

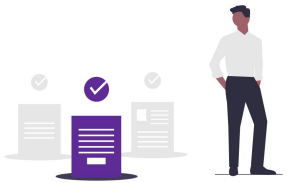
```
["In the sentance വൈകന്നേരം അവർ വീണ്ടും അവിടെ പോവുകയും ദൈവത്തെ ആരാധിക്കുകയും ചെയ്തു . the pronoun അവർ refers to ['വിനോദസഞ്ചാരികൾ', 'ഗ്രുവായൂർ']",
 "In the sentance വൈകന്നേരം അവർ വീണ്ടും അവിടെ പോവുകയും ദൈവത്തെ ആരാധിക്കുകയും ചെയ്തു . the pronoun അവിടെ refers to ['ക്ഷേത്രത്തിൽ']"]
```



Results & Discussion

Result & Application

- For simple sentence that is parsable by the underlying models, the naive algorithm performs well.
- When tested on random **30 wikipedia documents**, the algorithm performed with a **65% accuracy** - which is a good marking for one of the first works in Malayalam.
- The work can be applied to ***Malayalam chat applications***. The small conversation style text would be perfect.
- Question answering systems should work with well depending on the complexity of the documents we are extracting info from.



Discussion

- The algorithm is not abstract summarisation ready.
- The work's accuracy goes to 65% in large documents due reasons including the following
- The Chunking & Parsing and mlmorph algorithms though best in the use open source has low accuracy in real world uncurated corpus which is a direct bottleneck.
- The mlmorph is able to analyse only 45% of the words in an uncurated real world corpus.

Discussion

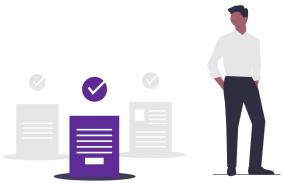
- Confusion of commonsense & context understanding . Depends on how you read it

പിണറായി വിജയനോട് പറഞ്ഞു

Did പിണറായി say to വിജയൻ

Or

did the person whose full name is പിണറായി വിജയൻ speak?



-

Comparison & BenchMarking

- This is the **first work in Anaphora Resolution in Malayalam** with ***verifiable results using public taggers and tools.***

- The other works are made from private datasets and taggers. The tools are not available publicly nor can it be recreated. Hence Benchmarking is not possible.

Future Work

- If the taggers which are bottlenecks to performance improves, the CR algorithm improves accordingly.
- Learnings from this can be transferred to Entity Resolution algorithms in Malayalam.
- If tagged corpus could be created, we can attempt machine learning methods or Deep Reinforcement learning like that in English which could improve performance.
- Study how such tools are developed could open new avenues of building NLP tools to resource poor similarly rich languages and could accelerate development

Thank you

Open for Questions

References

- Clark and Manning. 2016. Deep Reinforcement Learning for Mention-Ranking Coreference Models.
- Clark and Manning. 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations
- Clark and Manning. 2015. Entity-Centric Coreference Resolution with Model Stacking.
- Raghunathan, Lee, Rangarajan, Chambers, Surdeanu, Jurafsky, Manning. 2010. A Multi-Pass Sieve for Coreference Resolution EMNLP-2010,
- Recasens, Can, and Jurafsky. 2013. Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions.
- Recasens, Marneffe, and Potts. 2013. The Life and Death of Discourse Entities: Identifying Singleton Mentions.
- Lee, Recasens, Chang, Surdeanu, and Jurafsky. 2012. Joint Entity and Event Coreference Resolution across Documents.
- Lee, Peirsman, Chang, Chambers, Surdeanu, Jurafsky.Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task.

References

- http://www.lrec-conf.org/proceedings/lrec2016/topics.html
- Pronominal Anaphora Resolution using Salience Score for Malayalam, Athira S, Lekshmi T S, Rajeev R R, Elizabeth Sherly and Reghuraj P C, 2014 [IITMK, GE Palakad]
- VASISTH-An Anaphora Resolution System, Sobha [M.G.University],L B.N.Patnaik[Indian Institute of Technology], 2011
- Anaphora resolution in Malayalam and Hindi, Sobha Lalithadevi, Anna University, 2010
- A Generic Anaphora Resolution Engine for Indian Languages, Sobha Lalitha Devi, Vijay Sundar Ram,Pattabhi RK Rao, 2014
- A Survey on Anaphora Resolution Toolkits, Seema Mahato, Ani Thomas , Neelam Sahu
- Referential Entity Resolution in a Corpus of Malayalam, Dr.S.Rajendran,Tamil University