A hybrid approach to Tamil Morphological generation

Rajeswari Sridhar, Sugadev C, Mani Murugesan P, Vignesh N T
rajisridhar@gmail.com, sugadev.cse@gmail.com, manimurugesan1993@gmail.com, vignesh.n.t@gmail.com
Department of Computer Science and Engineering
College of Engineering, Guindy
Anna University, Chennai, India.

ABSTRACT

Tamil is regarded as one of the complex language as it is rich in words and highly agglutinative in nature. More than hundred inflections can be formed for a given word. This makes the processing of Tamil words highly difficult. Our work presents simple, efficient method to produce inflections of the Tamil verbs and nouns. Our work combines the advantages of the existing models by combining the rule based approach and the data driven approach.

General Terms

Data driven approach, rule based approach, natural language processing, morphological generator.

Keywords

Inflection rules, Suffix table, verb-forms, noun-forms, morpo-lexical inflection.

1. INTRODUCTION

Morphological generation is the process of producing the various inflection of the given root word based on the nature and rules of the pertaining language. It is the primary source for the translation of any language as it produces morpholexical inflection based on the given criteria (gender, tense, etc.). Since Tamil language is highly agglutinative in nature, each root is affixed with several morphemes to generate word forms. The challenge lies in choosing the right and appropriate word form. Our work combines the advantages of the existing systems and produces the inflections. Our morphological generator takes in the root word and the morpholexical inflection and produces the appropriate word form for the nouns and all inflections for the verbs.

2. RELATED WORKS

There are different morphological generators developed for various languages using different approaches.

For Indian language like Hindi, approaches like database based approach has been developed. This approach is simple because all the inflections are stored in the database and based on the given word the inflection can be directly found by querying the database. But it is highly difficult to pre-record all the morphological inflections of all the words of a morphologically rich language and storing in a database and querying it when needed.

There are other approaches like Ruled based approach [1] where the morphological inflections are produced by using the rules of the pertaining language. Here the inflection is generated by generating suffixes based on the tense, gender and other inflection criteria's by using the grammatical rules of the language. But matching the rules of a language and generating suffixes is a costly operation and it is considered slower than other operations.

One of the fastest approach is the data driven approach[2]. In this approach the suffixes are stored in the suffix table. Here the inflections are generated directly by affixing the suffix from the suffix table to the stemmed root word. This approach is efficient but the space required is larger than the other approaches. Our approach closely resembles the data driven approach.

3. MORPHOLOGICAL GENERATOR FOR TAMIL

Generally the method used for the morphological generation of the Tamil language is the rule based method but its computational time is little higher as it identifies the type and produces the inflection by forming the appropriate suffixes. Data driven approach is considered as fast as it produces inflection directly based on the suffix of the given word but it requires more pre-recorded data. Our work combines the advantages of the above two approaches and produces the inflecions.

Example for the inflexion of Tamil words.

Noun:கத்தி+instrumental=கத்தியால்

Verb:படி+past+singular+male=படித்தான்

3.1 Challenges in Tamil Morphological Generator

Tamil is morphologically rich and agglutinative in nature. Thus the foremost difficulty is in choosing the right inflection of the given word. There are various cases where choosing the correct inflexion is highly difficult. Example: போடு and தேடு

Both போடு andதேடு has the same suffix and the nedil as their prefix but they need different suffix in their inflexion

போட்டான் ,

தேட்டான்is not correct whereas தேடினான் is the correct inflection.

These kinds of exceptions are also handled in our generator by adding exceptional rules.

3.2 Suffix table

We have designed a suffix table is the essential part of the morphological generator as the words form inflections by accepting entries from the suffix table. Our generator requires less entries compared to other data driven generators.

Suffix table is a 2-D array where the row corresponds to the category and the column corresponds to the inflection. Table 1 is an example suffix table for nouns. A similar suffix table is constructed for the verbs.

Table 1: Suffix table for nouns and their inflections.

யை	யால்	யில்	யின்	யினது	…..
லை	லால்	லில்	லின்	லினது	…..

3.3 Algorithm

Our generator accepts the root word, the word class (noun or verb) and the inflection as the input and produces the output based on the following algorithm which is specified as a flow chart.

i. Segregate the word into individual letters using a defined regular expression.

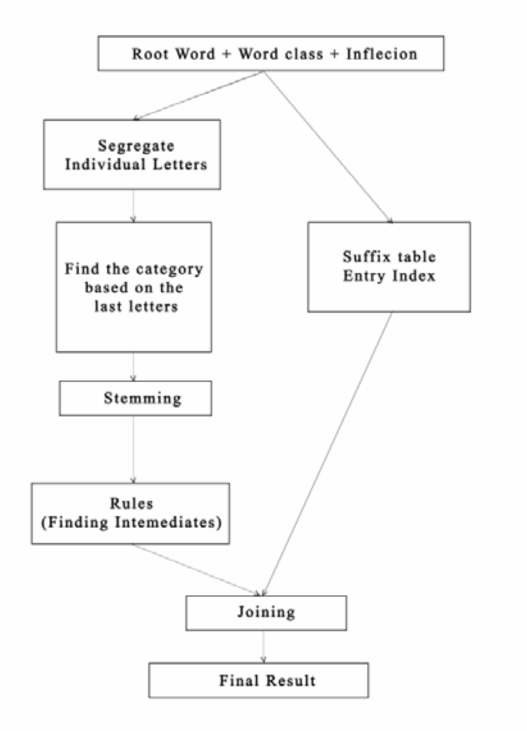
ii. Find the category based on the last characters

iii. Stem the unwanted characters.

iv. From the inflection specified by the user, find the suffix table entry index.

v. Find the intermediate that has to be added between the root word and the suffix.

vi. Then join the stemmed word, intermediate and the suffix table entry to form the result


Root Word + Word class + Inflecion
Segregate Individual Letters
Find the category based on the last letters
Suffix table Entry Index
Stemming
Rules (Finding Intemediates)
Joining
Final Result

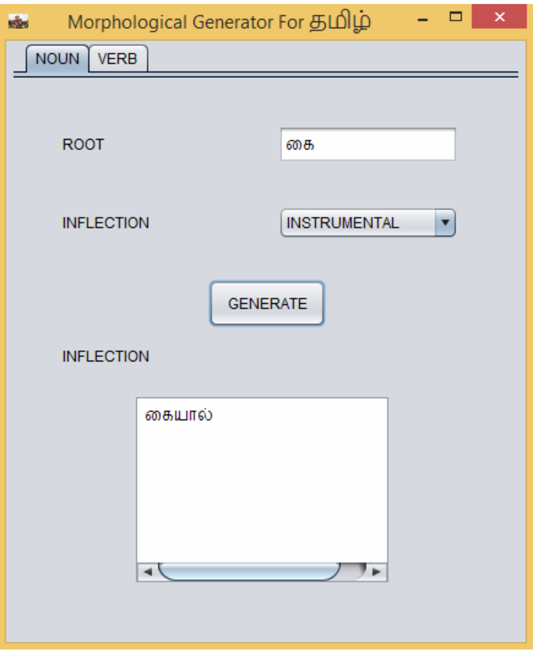
3.4 Features

- Our morphological generator can be used in the language translation systems.
- Uses very less stored data
- Combines the advantages of the existing systems
- Simple, efficient and high speed
- Can be easily updated
- New rules can be added easily.

4. GUI and Results

The GUI is developed using java. The GUI is simple that everyone can use it. The user has to just select the word class and enter the root word in the input field and then the inflection type. The output is displayed in the output field.

The following screen shots gives few of our output for the nouns and verb.

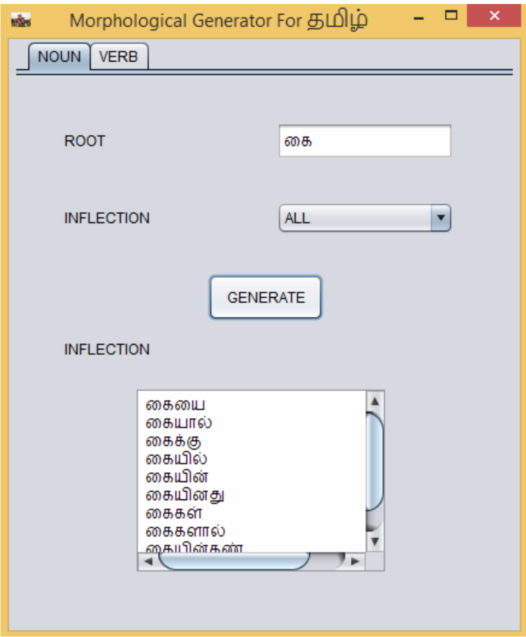


ROOT: கை

INFLECTION: INSTRUMENTAL

INFLECTION:

கையால்

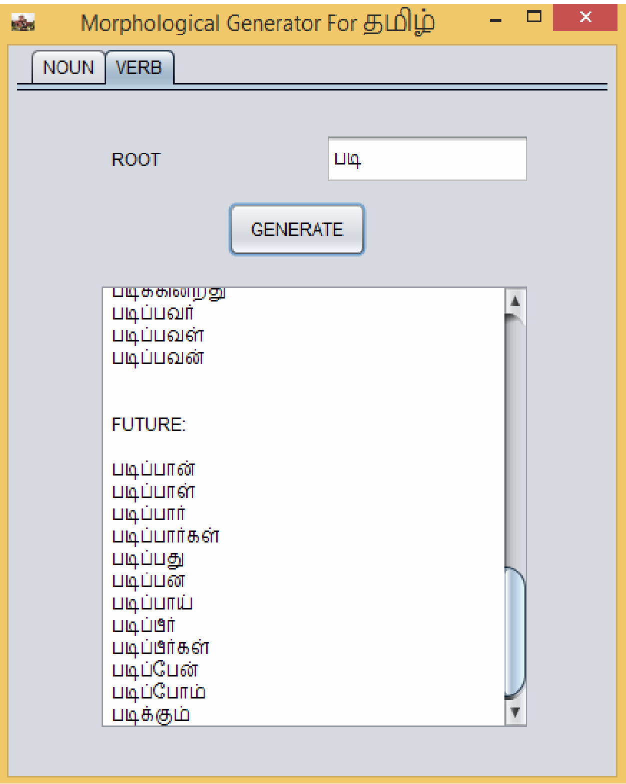


ROOT: கை

INFLECTION: ALL

INFLECTION:

கையை
கையால்
கைக்கு
கையில்
கையின்
கையினது
கைகள்
கைகளால்
கையின்கண்



ROOT படி

```
படிக்கின்றது
படிப்பவர்
படிப்பவள்
படிப்பவன்


FUTURE:

படிப்பான்
படிப்பாள்
படிப்பார்
படிப்பார்கள்
படிப்பது
படிப்பன
படிப்பாய்
படிப்பீர்
படிப்பீர்கள்
படிப்பேன்
படிப்போம்
படிக்கும்
```

The algorithm gave nearly 96% accuracy for nouns and verbs. The accuracy could also be increased by adding appropriate rules for failed scenarios. The sequence of the rules needs to be specified in order to not use an incorrect rule. However, the system has not been integrated with others and hence the efficiency in a combined environment could not be measured.

6. CONCLUSION AND FUTURE WORK

The algorithm explained here is a combination of all the advantages of the existing works. This method can be used for all morphologically rich languages. This algorithm can be used to generate morphological generator for pronouns and adverbs. This method can be used to create morphological generator for all the Dravidian languages. The integration of this system with other systems is yet to be tested.

REFERENCES

[1] P. Anandan, Dr. Ranjani Parthasarathy and T.V. Geetha, “Morphological Generator for Tamil”, INFITT, 2001

[2] Anand Kumar, M; Rekha, RU; KP Soman; Rajendran, S; Dhanalakshmi, V, “ A Novel Data Driven Algorithm for Tamil Morphological Generator”, International Journal of Computer Applications, Foundation of Computer Science, Volume 6, No. 12, p.52–56, 2010.

[3] http://en.wikipedia.org/wiki/Noun

[4] en.wikipedia.org/wiki/Grammatical_case