Tamil Wordnet based on hybrid approach

Rajeswari Sridhar , Jayavasanth R
rajisridhar@gmail.com,jairvasanth@gmail.com
Department of Computer Science and Engineering, Guindy,
Anna University, Chennai

Abstract: Wordnet is an attempt to find semantic relationships between words such as synonyms, antonyms, holonyms, homonyms, meronym etc., They play an important role in machine translation and various other NLP applications. This paper enumerates on the methodology for the construction of an independent wordnet for Tamil language based on a Hybrid incremental model.

1. Introduction

Tamil is a classical Dravidian language spoken predominantly by Tamil people of South India and North-east Sri Lanka. Tamil literature has flourished for over 2000 years in the past and has evolved a lot over the years with respect to the script and words.

2. Existing Tamil Wordnets

A Previous attempt [1] to construct a wordnet was based on synsets linking with English and Hindi synsets. The results of it, concludes that, synsets linking approach ended up in creating a Hindi-Tamil bilingual dictionary rather than a Wordnet and stressed on the importance of an Independent Wordnet for Tamil. In addition, their Wordnet comprised of only the synsets representation for the language ignoring all other semantic relationships.

3. Proposed Approach

Roughly, Wordnets are built using two approaches [2]: the merge model, where an independent subset is built and linked to another wordnet such as Princeton wordnet, manually and the expansion model where the synsets are taken from another wordnet and linked with the words of our language. The former involves manual effort while the latter disregards the semantic structure of our language and also may result in incorrect linking of subsets.

We aim to build an independent Tamil wordnet using incremental model by using a semi-automatic approach and is given in Figure 1.

Step 1: Input set

To generate the wordnet, words for EMILLE [3] corpus are used.

Step 1.1: Morphological Analysis of words

In Tamil, each word can form a few hundred words by addition of various suffixes, which indicate the tense, gender, plural or singular etc.,

For e.g.

ஆடு – ஆடினான் , ஆடினாள் , ஆடுவான் , ஆடுவாள் , ஆடுவார்கள் , ஆடுகின்றனர்.

The root word of all these terms is 'ஆடு'.

In case of 'ஆடினாள்'

ஆடினாள் – ஆடு + இன் + ஆள்

The suffix'இன்'denotes that the event has occurred in the past and 'ஆள்' denotes that the doer of the action is a female. We do consider only the root word while forming the synsets and the various morphological derivations of the root word are marked and stored along with the root for future reference.

To extract the root word from the corpus a Morphological analyzer developed by Anna University's Tamil Computing Laboratory (TACOLA) [4]. Also it identifies the Part Of Speech of the word.

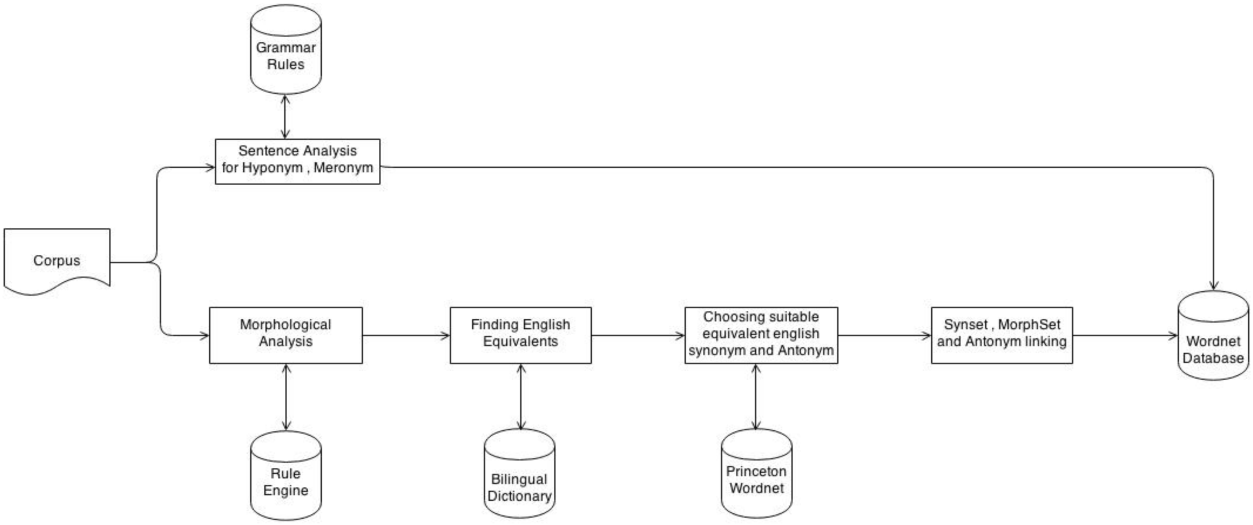
Grammar Rules
Sentence Analysis for Hyponym , Meronym
Corpus
Morphological Analysis
Finding English Equivalents
Choosing suitable equivalent english synonym and Antonym
Synset , MorphSet and Antonym linking
Wordnet Database
Rule Engine
Bilingual Dictionary
Princeton Wordnet

Fig 1.Block Diagram

Step 1.2: Finding English Equivalents

Rather than linking with synsets, we intend to find equivalent English words for the root word using a bilingual dictionary.

Step 1.3: Choosing suitable English Synonym and Antonym

Among the words found as such, the shortest word which describes the Tamil word best is chosen as its English Equivalent. Also the POS of this equivalent English word needs to be checked using Princeton Wordnet [5].
e.g.: படிwill give both stair and read in any bilingual dictionary. We should not mistake 'படி' in 'படித்தான்' for a noun (stair).This word will be used to link the subsets. The English equivalent is checked with the Princeton wordnet to fetch an Antonym for the same.

Step1. 4: Synsets, Morph set and Antonym linking

A graph structured database - neo4j [6] is used at the backend for easy access and to mimic the way of linking as in nets.

There are two types of nodes: word-nodes and sentence-nodes. Each sentence node will represent a sentence from the corpus whose attributes will the sentence itself. Word-nodes can either be a root word-node or a morph word-node. Each root-word-node will represent a Tamil word whose attributes will include the English equivalent and its POS. And each morph word-node will carry information regarding its suffixes.

For every corpus sentence a sentence-node is created once and once for every word in it (which has not occurred so far) a root word-node and morph word-node (if different from the root) are created once with required data. The sentence-node is linked to the morph node using “used in” relation (or edge), which in turn is related to the root node using “morph” relation.

The root word node based on its English Equivalent is linked to one of the root word nodes having the same English Equivalent and POS using “synonym” relation. Also the new root word node is linked to the one of the root word node having its antonym as the equivalent English word using “antonym” relation.

In this manner, a set of root words that are connected using a “synonym” relation forms the synonym set.

A structure of this graph database can be visualized as in Fig 2.

Step 2: Sentence Analysis for hyponyms and meronym

During this step , more information from the input corpora is obtained by analyzing the sentence and looking for indicators such as எனும் , என்கிற , எனப்படுகின்ற , ஒரு வகையான , ஒரு ரகமான , இன்.,

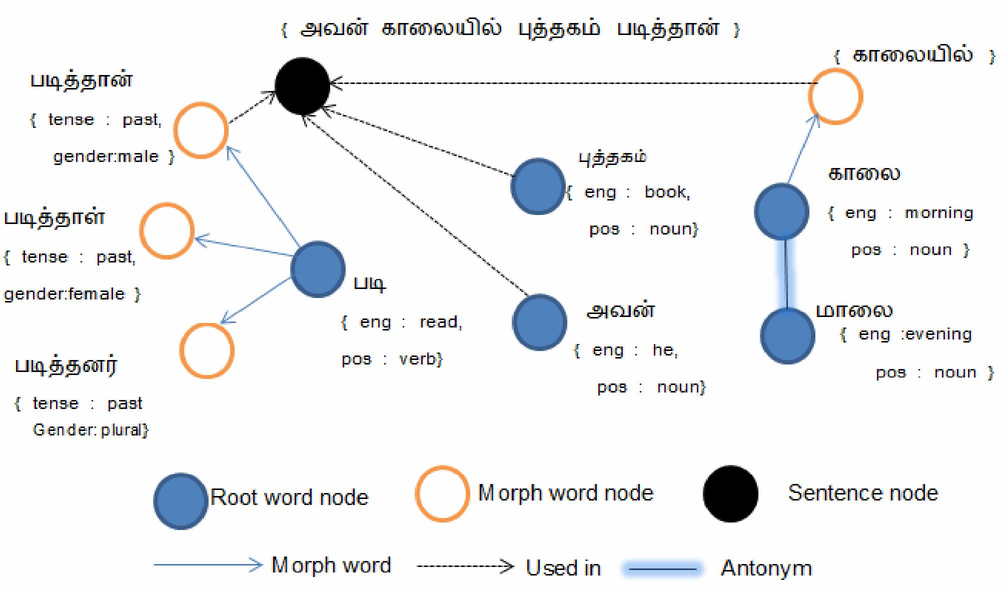

Fig 2. Visual representation of the database

These indicators specify the occurrence of hyponyms and meronym based on the noun words on either side of them to a greater extent.

For e.g.

அன்னம் என்னும் ஒரு பறவை நீரையும் பாலையும் பிரித்தெடுக்கும்.

“Annam ennum oru paravai neerayum paalayum pirithedukkum

Here the word ‘என்னும் ஒரு’ acts as an identifier , thereby specifying that ‘அன்னம்’ is a hypernym of ‘பறவை’ and is shown in Figure 3.

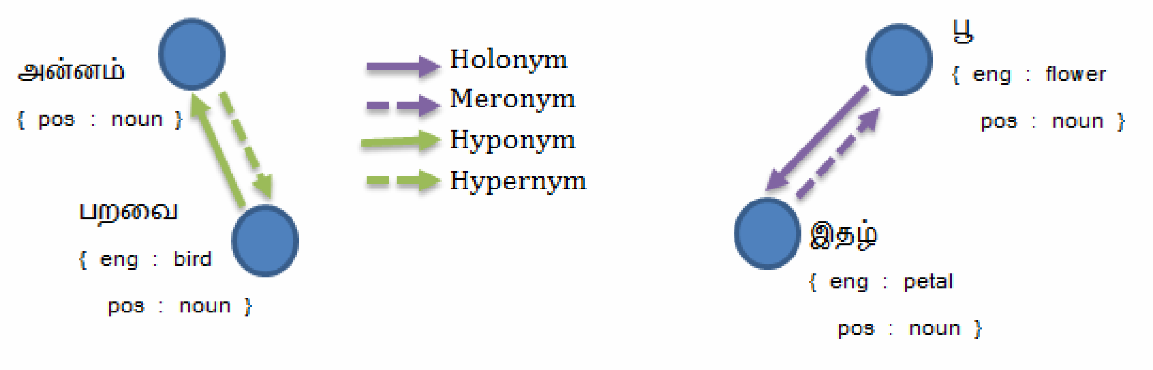

Fig 3. Hyponyms and Meronym – rep. in database

Also, பூவின் இதழ்கள் வாடி விட்டன

“Poovin ithalgal vaadi vittana”

Here the suffix ‘இன்’ of ‘பூவின்’ will act as an identifier to classify ‘இதழ்’– ‘பூ’under meronym – holonyms set.

Sample Input

“அவன் காலையில் புத்தகம் படித்தான்”

Output of the Analyzer

அவன் - Noun

காலையில் - காலை + ய் + இல் – Noun

புத்தகம் - Noun

படித்தான் - படி + த் + த் + ஆன் - Verb

Output of the wordnet

The output will be as shown in Fig 2.

5. Results Analysis

Nearly 2146 words have been analyzed and about 460 relationships have been found between them. Implementation of the word-net as Graph database has certainly proved better.

But this approach works well only if the same English equivalents are present for two words in the bilingual dictionary. If it is not found, linking has to be done in the background. The number of words in the bilingual dictionary limits this approach. Also the errors in Morphological analyzer are percolated in this process. This independent wordnet has to be expanded further using different other corpuses.

6. Conclusion and Future Work

This work is based over a single bilingual dictionary that puts a limit on the set of words that can be used. Compiling a bilingual dictionary out of all existing ones could prove better in the case where a variety of corpuses are being used for input. Also, Sentence Analysis to find the sentence structure could help in cases where the Morphological Analyzer fails. The integration of this wordnet with other systems needs to be analyzed for its ability to scale with other systems.

References

[1] Rajendran, S., Arulmozi, S., Shanmugam, B., Baskaran, S., Thiagarajan, S.: Tamil WordNet. In: Proceedings of the First International Global WordNet Conference, Mysore, vol. 152, pp. 271–274, 2002.

[2] Enabling Minority Language Engineering (www.emille.lancs.ac.uk)

[3] Oliver, A., Climent, S.: Building wordnets by machine translation of sense tagged corpora. In: Proceedings of the Global WordNet Conference, Matsue, Japan (2012)

[4] Anandan, P., Saravanan, K., Parthasarathi, R., &Geetha, T. V., : Morphological analyzer for Tamil : Proceedings of ICON2002, pp. 3-10, 2002.

[5] Princeton University "About WordNet." WordNet.Princeton University. 2010.
(www. wordnet.princeton.edu)

[6] Neo4j is an open-source graph database supported by Neo Technology (ww.neo4j.com)