Enhancing Tamil WordNet with Subcategorization Information

Vijay Sundar Ram R and Sobha Lalitha Devi
AU-KBC Research Centre
MIT Campus of Anna University, Chennai-600044
sobha@au-kbc.org

Abstract

We discuss about enhancement of Tamil WordNet, with subcategorization and selectional restriction rules as features for nouns and verbs respectively. This subcategorization is from language based ontology. This helps in finding the verb and its arguments, which is needed for many NLP applications.

1 Introduction

WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. Inline with WordNet by George A Miller, Wordnet in different languages were developed. Here we discuss about an existing WordNet for Tamil and enhancing present word net by adding subcategorization as a feature. WordNet is used in different NLP applications such as machine translation, in word sense disambiguation, even in information retrieval to enhance the query.

2 Tamil WordNet

Tamil WordNet is an attempt to build a lexical network for Tamil language along the lines with Princeton WordNet (Fellbaum 1998). In Tamil WordNet, there are information about nouns, verbs, adjectives, and adverbs. These are organized in the notion of a synset. In this was attempted to assign all possible word-level relations such as synonymy, hypernymy, hyponymy, meronymy, holonymy and antonymy. This contains semantic relations for 50,000 words in Tamil along with their sense. In nouns, the relations such as synonymy, hyponymy, and meronymy are captured. In case of nominal forms corresponding verb forms are represented as relations. In verbs, relations such as synonymy, hypernymy, troponymy, nominal forms, related nouns are captured. In adjectives only few root adjectives are listed. There exist derived adverbs and derived adjectives. In these the synonymy and related noun relations are handled.

The total number of words available in each part-of-speech category is present in the table 1.

Table .1 Statistics of Words

Part-of-Speech	Total Words	Unique Senses
Noun	46710	37530
Verb	2881	2423
Adjective	416	388
Adverb	490	423

3 Tamil Wordnet with subcategorisation

Verb is the nucleus of a sentence and it has the right to select its arguments. This is called the selectional restriction constraint for the verb. These Selectional Restrictions (SR) in the language with respect to the units of a sentence, were used for generating meaningful sentences. Subcategorization features such as ±concrete, ±animate explain the nature of the noun. The verb in a particular sense can take the arguments only according to its selectional restriction constraints. Here we add the selectional restriction rules as a feature for each verb and the subcategorization as a feature in the nouns (Arulmozhi 2006).

Subcategorization features explain the nature of the noun. Essentially, the subject nouns and object nouns are analysed using these features. These features may include the type of noun, its characteristics, state etc. Subcategorization information includes the features such as [±animate], [±concrete], [±edible] etc. Some of the features and the meanings are listed below.

[+animate] - all animals, human beings
[+human] - all human beings
[+solid] - things which are in solid state
[+vehicle] - all the vehicles
[+concrete] - things that are physically existing
[-concrete] - things that do not physically exist
[+edible] - things that can be eaten
[+movable] - things that are movable
[-movable] - things that are not movable
[+avion] - things that can fly.
[-avion] - things that can not fly.

When these features are assigned to nouns, in a sentence, we get more semantic information about the noun in that sentence. Very fine features such as [+possession], [+furniture] are also assigned. Some examples of nouns and their subcategorization features are given below.

The subcategorization features of the noun "kaar"(car) give the characteristics of car. It is a non-living entity, physically existing, solid, manmade object. This is a vehicle, which has wheels. The features below give these characteristics.

kaar (car):
[-living, +concrete, +movable, +artifact, +solid, +instrument, +vehicle, +wheeled]

The subcategorization features of the abstract noun "katinam" (hard) are given below. This is not a physically existing entity, and this is not virtual. This is a feeling which can be sensed.

katinam (hard) :
[-living, -concrete, -virtual, -feature, +sensible, +feeling]

Some nouns such as "miin" (fish) can have more than one set of features. It can be a living being as well as a food item. The following example illustrates that.

miin (Fish)
[+living, +animate, +vertebrate, -mammal, -avion,**+fish**]
[-living, +concrete, +movable, +food_items, +solid, **+animal_prod**]

These subcategorization features are the nodes of the language based ontology, which is represented in a tree structure.

The commonly used ontologies are more based on the taxonomy of nature (Noy, 2001) (York Sure, 2002). For example, the most commonly used lexical resource, Wordnet is based on natural classification. It gives all the senses of a word (Miller 2001). It defines the relationships such as hyponymy, synonymy etc., but it does not define the relationship between the verb and its arguments in a sentence and it does not give the subcategorization features too. A language based ontology gives the relationship between the verb and its arguments and it is needed for NLP activities because, it can provide more information without doing deep parsing, which is expensive.

3.1 Features of the Ontology

The ontology discussed here is developed with the perspective of how the nouns could co-occur with verbs in English sentences. Though it reflects some features in nature, it substantially deviates from the taxonomy of nature.

In general the ontology has the following features.

1. It is a language based ontology.
2. Nodes in the ontology are the subcategorization features of the nouns.
3. This is different from the taxonomy of nature.
4. The hierarchy is made according to the usage of nouns in the language.
5. Each node will have a list of nouns as entries of that node.

The ontology is formed using the subcategorization features of the noun. These features are organized in to a tree structure. These subcategorization features represent the type and nature of the nouns.

The ontology starts with a root node "entity" under which any noun can be placed. We define that any noun is an entity. Then the root node is further classified into living and non living entities. They are represented as [+living] and [-living]. The [+living] node is divided into animate and inanimate (±animate). The [+animate] node is further classified into ±vertebrate. This is further classified into ±mammal, where +mammal is subdivided into ±human as sub nodes and –mammal as ±avion. The +human sub node is further classified into ±female, which is the leaf node of the tree [+living, +animate, +vertebrate, +mammal, +human, ±female] and is the sixth level of the hierarchy. There are 3 important major nodes when [-living] is subdivided.

Concrete, abstract and virtual entities. [+concrete] specifies the things which have physical existence (Ex: "meejai" (table)). [-concrete] specifies the things which do not have physical existence (Ex: "ennam"(thought)). +virtual specifies the things that do not have a physical existence, which undergo some of the actions that the physically existing entities (Ex: "thiraipadam"(movie)).

3.2 Difference from Wordnet

In Wordnet, all the nouns are classified as either a physical entity or abstract entity. But for the purpose of subcategorization and SR of verbs, there should be another classification virtual entities. According to Wordnet, e-mail is an abstract entity. Internet and World Wide Web are physical entities. All these are classified under virtual in our ontology.

Take the example "kaditham" (letter), this noun is classified as an abstract entity in the Wordnet. But this has to be a physical entity for certain verbs to satisfy the subcategorization rule. Consider the example:

```
Noun : kaditham
Sense: a written message ad-
dressed to a person
Usage (Example Sentence):
'avan kadithaththai
he    letter+acc
kiziththaan'
tear+past+3SM
"He tore the letter"

Ontology Entry:
[-living, +concrete,
+movable, +artifact, +solid,
-instrument, +creations]
```

In the above example, the verb "kizi" (tear) can take only a [+concrete] as an object. But, Wordnet classifies "letter" as an abstract entity. So, in our ontology, "letter" is [+concrete].

The nouns, such as "makkaL"(people), is classified under abstract entity in the Wordnet as in the following sentence

```
"makkaL kalanthukkolla
  People     to join
mutivetuththanar"
decide+PAST+3PL
(People decided to join.)
```

The verb "mutivetutu"(decide) can have the subcategorization [+living, +animate, +vertibrate, +mammal, +human]. But according to Wordnet, "makkaL"(people) is an abstract entity. This violates the subcategorization rule. So, in ontology, the noun "makkaL" is an entry in the node [+living, +animate, +vertibrate, +mammal, +human].

```
*"arivu   kalanthukkolla
  knowledge to join
mutivetuththu."
decide+PAST+3N"
```

The above sentence is a semantically wrong sentence, because, "arrivu" (knowledge) is an abstract entity. It cannot join anything.

In nature, things are classified in the perspective of how it is viewed or perceived. For example, the noun, "caalai" (road) is a non-movable entity which cannot move from one place to another place. Consider the following example

```
"intha caalai engaL uurkku
  this road   our town+dat
celkirathu"
go+PRE+3PN
(This road goes to our
town.)
```

The above example is a meaningful sentence both syntactically and semantically. The verb "cel"(go) can take only a "movable" entity as a subject if you consider the semantic properties of "cel". But it can take non-movable nouns as its subject. In nature, "caalai"(road) is a non-movable entity, but it takes "cel" (go) as the verb. Thus in the ontology we have to place road under a different feature [+movable]. This is where the ontology based on language makes the difference from nature.

This distinguishes our Ontology from Wordnet hierarchy. If Wordnet is used, validating the wrong sentences of this type is not possible.

Here we are not considered figurative usage while developing the subcategorization.

Consider the example

```
"thalaivar meedaiyil
Leader    in the stage
kargithaar."
roar+PAST+3SM
(Leader roared in the stage)
```

Here the verb "kargi"(roar) will take "cingkam"(lion) is it subject, whose subcategorization

is ([+living, +animate, +vertebrate, +mammal, -human, -avion, +carnivorous, +dog_family]). But as a figurative usage the above sentence is a valid sentence.

The difficulty of placing a noun in ontology increases when it is an abstract noun. For example the noun "civappu"(red) is a color, which could be attributed to a physically existing thing. But identifying it as a Psychological feature or as a physical feature is difficult. The use of this noun with the verb "maariyathu"(turned) has shown that it could be assigned to [+physical_feature]. As in the sentence:

"avan mukam koopaththil
His face in anger
civappaaka maariyathu."
red+adv turn+PAST+3SPN
(His face became red with
anger)

Hence, it comes under,
[-living,-concrete,-virtual,+feature, +physical_feature].

Nouns such as "thiraipatam" (movie), "e-meyil" (e-mail) etc. can be easily classified under virtual entities. The nouns such as "pangku canthai" (stock market) are difficult to identify whether it is a physical or abstract entity. They come under [+virtual] in the ontology.

In the ontology, [+plants] do not have the classification of [±movable]. But some water plants which are not fixed to the land with roots have the nature of moving. That kind of verbs, which take only moving and do not take non-moving plants (or vice-versa) as arguments, we have not come across in the analysis. So, there is no classification of [±movable] for [+plants].

3.3 SR RULES

Rules, which analyse a category in terms of syntactic features, are called SR rules. The SR rules are developed according to the type of verb and the number of arguments that the verb can take. The SR rules are made according to the verb and the co-occurrence of subject and object. If a verb takes certain kind of subject, there can be a constraint for the object that only particular kind of object can occur (Sobha 1989). Consider the examples given below.

"avaL aappil caappittaal"
she apple eat+PAST+3SF
(She ate an apple.)

The verb in sentence is "caappitu"(eat). This is a dyadic verb, which takes two arguments, the subject and the object. The subject of the sentence is the one who does the action "caappitu"(eat). Only a living being can do the action "caappitu"(eat). So, the subcategorization rule for the subject should be [+animate]. The verb "caappitu"(eat) can take the objects, which are solid edible items. The items, which are not in solid state cannot become an object for the verb "caappitu". The syntactic subcategorization rule for "caappitu"(eat) is represented as follows.

RULE :

Verb : "caappitu"(eat)
Type : Dyadic
Syntactic Arguments:
Subject: [+living,+animate]
Object: [-living, +concrete, +movable, +food_items, +solid]
[-living, +concrete, +movable, ±artifact, +solid, +edible]

In the above rule, there are two rules given as object rules. One is the categorization for solid food items such as cake, apple etc. another object rule is for other natural or artificial solid edible items such as tablet etc. Verbs such as chew, swallow, munch etc. takes the same rule.

Consider the verb "ootu" (run). The syntactic subcategorization rule for "ootu" (run) is represented as follows.

RULE :
Verb : "ootu" (run)
Type : Monadic
Syntactic Arguments:
Subject:
[-living, -concrete, +virtual, -comp]
[-living, +concrete, +movable, +artifact, +solid, +instrument, +vehicle]
[-living, +concrete, +movable, -artifact, +liquid]
[+living, +animate]
Object: [No Object]

Here the verb "ootu" can take subject arguments with different subcategorization. The example sentences are given below

"pirokitam ootikkontuirun-
thathu."
(The program was running)

Here “pirokitam” (program) has the subcategorization [-living, -concrete, +virtual, -comp].

“kaar intha vaziyil ootiyathu.”
(The car ran through this way)

Here “kaar” (car) has the subcategorization [-living, +concrete, +movable, +artifact, +solid, +instrument, +vehicle]

“niir intha kulaayil ootiyathu.”
(Water runs through this tube)

Here “niir” (water) has the subcategorization [-living, +concrete, +movable, -artifact, +liquid]

“naaykaL kaattilirunthu ootina”
(Dogs ran from the forest)

Here “naaykaL” (dogs) has the subcategorization [+living, +animate]

In English the verb “run” occurs as monadic and dyadic verb. And take different subcategorization nouns as subjects in both the cases (Arulmozhi 2006).

In this study, we have taken 2600 verbs, for analysis. The selection of verbs was according to Wordnet concepts and most frequently occurring verbs. The synonyms of the verbs are also considered. We have confined our analysis to the most common sense of all the verbs. The verbs were grouped into 184 groups according to the subject and object it takes. In those groups, if a commonly used sense of a verb is not present, that is included as an exception. Further we classified the verbs according to the subject it takes. We have 47 classes of verbs. In this, a verb can occur in more than one class, if it takes more than one rule for the subject. The table below gives the part of spread of verbs in those 47 classes.

Table .2 Verb Class

S.No	Subject Rules	No. of Verbs	Verb Class
1	[+living, +animate, +vertebrate, +mammal, +human]	2489	Human Verbs
2	[+living, +animate]	658	Animate Verbs
3	[-living, +concrete, +movable, +artifact, +solid, +instrument]	268	Instrument Verbs
4	[-living, +concrete, +movable, +artifact, +solid]	244	Solid Artifact Verbs
5	[-living, -concrete, -virtual, -feature, +event]	188	Event Verbs
6	[-living, -concrete, -virtual, -feature, +content]	171	Content Verbs
7	[-living, +concrete, +movable, +artifact, +solid, +instrument, +vehicle]	110	Vehicle Verbs
8	[-living, +concrete, +movable]	76	Movable Verbs
9	[-living, +concrete, +movable, -artifact, +solid]	67	Movable Natural Solid Verbs
10	[-living, -concrete, +virtual, -comp]	62	Virtual non-computer Verbs
11	[-living, +concrete]	60	Concrete Verbs
12	[-living, +concrete, -movable, +artifact, +buildings]	53	Building Verbs
13	[-living, +concrete, +movable, +food_items, +solid]	49	Solid Food Verbs
14	[-living, +concrete, -movable, +artifact]	40	Immovable Artifact Verbs
15	[-living, +concrete, +movable, -artifact, +liquid]	38	Natural Liquid Verbs
16	[-living, +concrete, +movable, +artifact, +liquid]	36	Artificial Liquid Verbs
17	[-living, +concrete, -movable, -artifact, +natural_object]	33	Immovable Natural Object Verbs

18	[-living, +concrete, +movable, +food_items, +liquid]	27	Liquid Food Verbs
19	[-living, +concrete, -movable, +artifact, +roads]	26	Road Verbs
20	[-living, -concrete, -virtual, -feature, +phenomenon]	24	Phenomenon Verbs

From the table, we could see that the most commonly used verbs are more oriented towards the human activity. The nodes, which are deep in the ontology, nearing the leaf nodes take more number of verbs than the top nodes such as [+living]. When it goes very deep such as [+domestic], [+creation], the number of verbs reduces. The only exception here is [+living, +animate] which is a top node, and takes more number of verbs.

This kind of verb classification is used for finding out the subject and the object of a verb. And after finding them, the semantic role also can be assigned to the subject and the object of a verb in a sentence. These rules work well for commonly used senses of verbs. For the verbs, which take very broad subcategorization rules, there can be some violations at the finer level, where it may become a figurative usage.

4 Conclusion

Here we have discussed the enhancements we made to the Tamil WordNet. We have classified the verbs based on the subject subcategorization it takes. We have added subcategorization features to the nouns and selectional restriction rules to verbs. The subcategorization features are obtained from a language based ontology, which is different from the English WordNet.

References

Arulmozhi, P. 2006. *Symantic Tagging for Language Processing*, Unpublished MS (By Research) Thesis, Anna University Chennai, TamilNadu, India.

Christiane Fellbaum (1998) 'WordNet - An Electronic Lexical Database'. MIT Press, Cambridge, London.

George A. Miller, et.al. 2001. *WordNet*, Cognitive Science Laboratory Princeton University. URL: http://wordnet.princeton.edu/

Gruber, T. R. 1993. *A Translation Approach to Portable Ontology Specifications' Knowledge Acquisition*, Vol - 5(2) pp.199-220.

John F. Sowa , *Building, Sharing, and Merging Ontologies* URL: http://www.jfsowa.com/ontology/guided.htm

Natalya F. Noy and Deborah L. McGuinness. 2001. *Ontology Development 101: A Guide to Creating Your First Ontology' Technical Report KSL-01-05*, Stanford Knowledge Systems Laboratory, Stanford University, Stanford, CA.

Sobha, L. 1989. *Selectional Restrictions In Malayalam - A Computer Aided Study*, Unpublished MA Dissertation, Department of Linguistics, University of Kerala, India.

York Sure, Michael Erdmann, Juergen Angele, Steffen Staab, Rudi Studer, Dirk Wenke. 2002. *OntoEdit: Collaborative Ontology Development for the SemanticWeb*, Proceedings of the first International Semantic Web Conference 2002 (ISWC 2002), June 9-12 2002, Sardinia, Italia.