Dr. R. Padmamala

WORD LEVEL TRANSLATION (TAMIL – ENGLISH) WITH WORD SENSE DISAMBIGUATION IN TAMIL USING ONTNET.

Dr.R.Padmamala[*1]

[*1]Asst. Professor, Dept of MCA, Ethiraj College for Women, Chennai, Tamilnadu.

Abstract: This paper discusses a knowledge-engineering approach for word level translation of Tamil words to English. This approach also shows the need for word sense disambiguation to arrive at the contextual category or meaning of a word when different category/meaning can be assigned to a particular word in different contexts. An ambiguous word in a given Tamil sentence is subjected to contextual analysis. The VPs in the given text are located and then augmented with semantic information. These semantic features are captured using the ontology derived from the sub-categorization features. This semantic information will help in assigning the correct meaning for the given ambiguous word. A rule-based syntactic parser and word sense disambiguator have been developed. This has been tested on Tamil sentences from Tamil newspaper websites and the results are encouraging.

INTRODUCTION

A Word Level Translation system is the core element of a Machine Translation system. A word in a language may have more than one meaning. Figure 1 illustrates this point in case of Tamil language. In turn, each of these words may have many translations back. This paper aims at ascertaining one meaning to a Tamil word based on the context in which it is present.

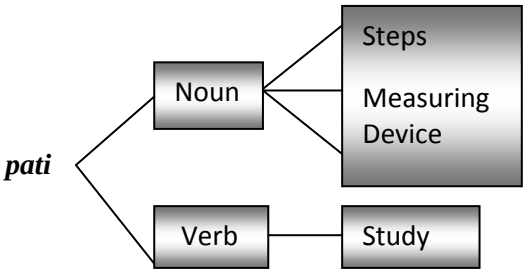


Figure 1

A word can be inflected or non-inflected. When it is considered as a non-inflected, root word, at that time a meaning is possible. If it is considered to be inflected, then it is morphologically parsed and then the meaning is ascertained. Figure 2 illustrates this point.

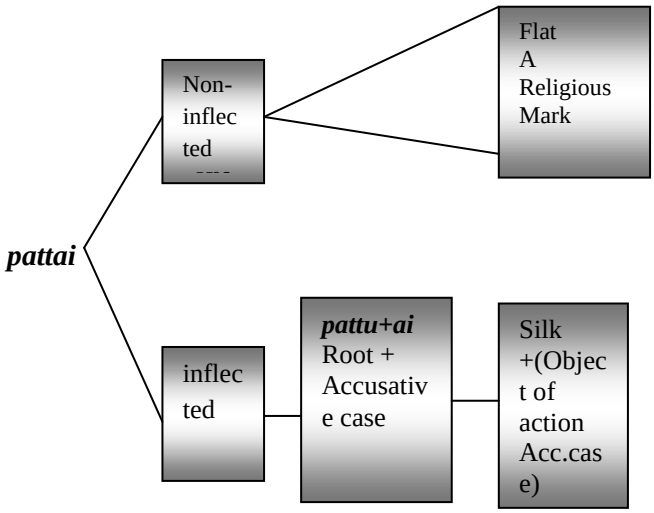


Figure 2

Also, for some inflected words, more than one way of morphological parsing is possible. Thus the meaning will also differ. This can be explained with Figure 3.

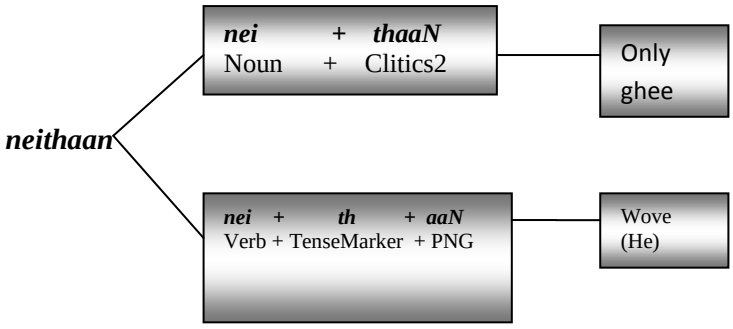


Figure 3

The purpose of this paper is to: 1. Analyse how contextual meaning can be elicited using semantic features 2. Examine how to combine the concepts of Ontology and Word Net for word sense disambiguation and 3.To substantiate the improvement in efficiency of word level translation when the combination of above said concepts is used.

The paper has the following schema. Syntactic parser is presented in the following section. Word sense disambiguator and word level translator incorporating ontology and word net are discussed in detail in further sections.

Existing system for Tamil – English Translation

2.1 Google Translator

	Input Sentence	Google Translation	Correct Translation
1.	*puthakaththaip pati*	According to the book	Read the book
2.	*avaN pattai neithaaN*	He ghee Bar	He wove the silk
3.	*veelai vaNnanku*	Work Worship	Worship the spear
4.	*kaththi azuthaaN*	Cried knife	Screamed and cried (he)
5.	*ati vaanku*	Step to buy	Get beatings
6.	*Ati kotu*	Give feet	Beat
7.	*naNRaaka ootu*	Nice tile	Run nicely
8.	*naLiNamaaka aatu*	The elegance of the goat	Dance elegantly
9.	*paathiraththai vazi*	Role in the way	Empty the vessel
10	*paathiraththai kazuvu*	Wash role	Wash the vessel
11	*oru piti sooRu*	A BT Rice	A handful of rice

Table 1: Google Translator

Clearly the output shows that the contextual meanings of the Tamil words are not captured. Therefore there are errors in the translation. This paper explains how all these errors can be corrected with rule-based method. The methodology followed is explained in the subsequent section.

METHEDOLOGY

In rule-based word level translation, the following steps are followed.

- The given sentence is tokenized.
- The constituent words are subjected to shallow morphological parsing, morpheme-labelling and word-class tagging.
- Person, number, gender, tense markers and case markers are retrieved.
- The constituent words are tagged with their corresponding POS tags.

These features give semantic information about the words. Using this information, word sense disambiguation can be done efficiently and more appropriate translation can be given to ambiguous words

1. Word sense disambiguator and translator

Word sense disambiguation is vital for areas such as machine translation, summarization and question-answering system and so on. In this paper, a combination of word net and OntNet is used to disambiguate a given word. Sub categorization features are added to the noun and verb involved using WordNet and OntNet.

1.1 OntNet – Establishing relationship between Noun Ontology and Verb Ontology

OntNet is an additional semantic network that is established between the Ontology tree of Noun entities and Ontology tree of Verb entities. In a WordNet, a word's semantic relation as a hypernym (superordinate), hyponym (subordinate), synonym, antonym to other words will be depicted. Mostly these relations are established between words of same category i.e. a synonym or antonym of a noun will

also be a noun. A semantic relationship that can exist between a word belonging to the Noun category and that to a Verb category is not established in case of WordNet.

OntNet focuses on establishing the semantic relationship between Noun ontology

tree and Verb Ontology tree. An OntNet of Tamil is formed by a collection of ontsets. Usually a synset in WordNet constitutes synonymous words. But in the OntNet that will be developed for Tamil, words of same semantic category can be grouped to form a ontset.

For example, in IndoWordNet (http://tdil-dc.in/indowordnet) the word ***vipuuthi*** **'Sacred ash** is grouped with the words ***thiruniiRu*** and ***thuNNuuRu*** which have the same meaning. These three words together constitute a synset.

Synset ID :	9635 **POS** : NOUN
Synonyms :	விபூதி, திருநீறு, துன்னூறு, தின்னூறு,
Gloss :	ஒன்றை தலை மற்றும் புஜங்களின் மீது பூசிக்கொள்ளும் ஏதாவது ஒரு ஹோமத்திலிருந்து துறவி மூலமாக கொடுக்கப்பட்ட ஒரு பொருள்
Example statement :	"மகாத்மாஜி நோய்வாய்ப்பட்ட குழந்தையின் உடலில் விபூதி இடுகிறார்"

Figure 4. Example for Synset

But in an OntNet, an OntSet is formed by grouping words that may have different meanings but fall under same semantic category. The following is an example of a Noun ontset.

OntSet ID 1: NOUN

veel, vil, thiruneeRu, kunkumam, thirisuulam, canku, cakkaram– Religious materials.

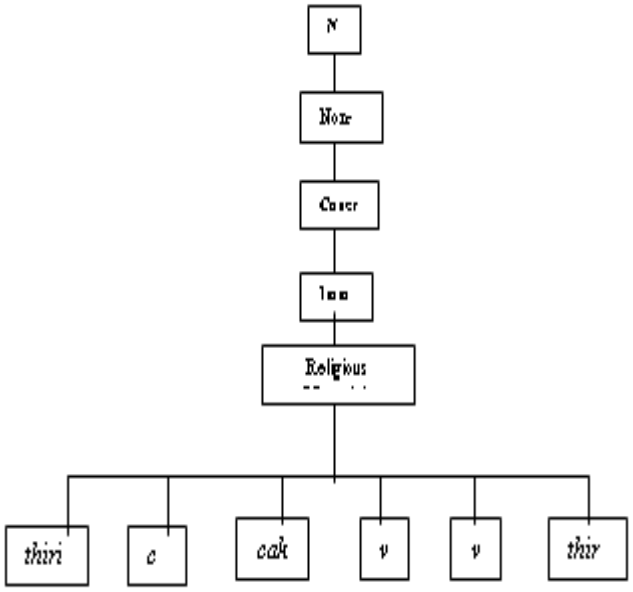


Figure 5. Example for Noun OntSet

OntSet ID 2: VERB

vaNnanku, kumpitu, thozu, thuthi **- Different forms of worship**

(*iRaivaNai vaNnanku* 'Worship the Lord**,**

vizunthu kumpitu 'procastrinate yourself in front of the Lord**,**

thiruvatikaLaith thozu 'Worship the Lord's feet**,**

iRainaamaththaith thuthi 'Worship the name of the Lord**)**

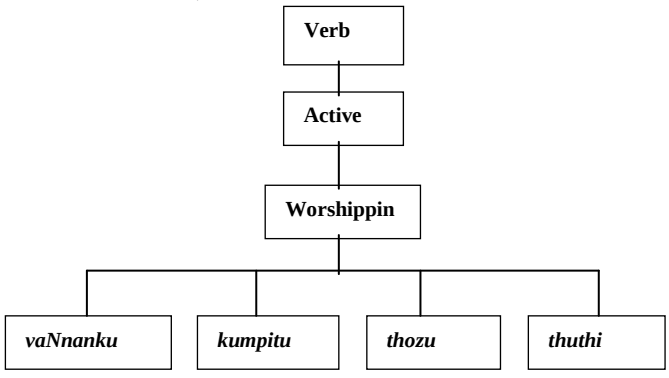
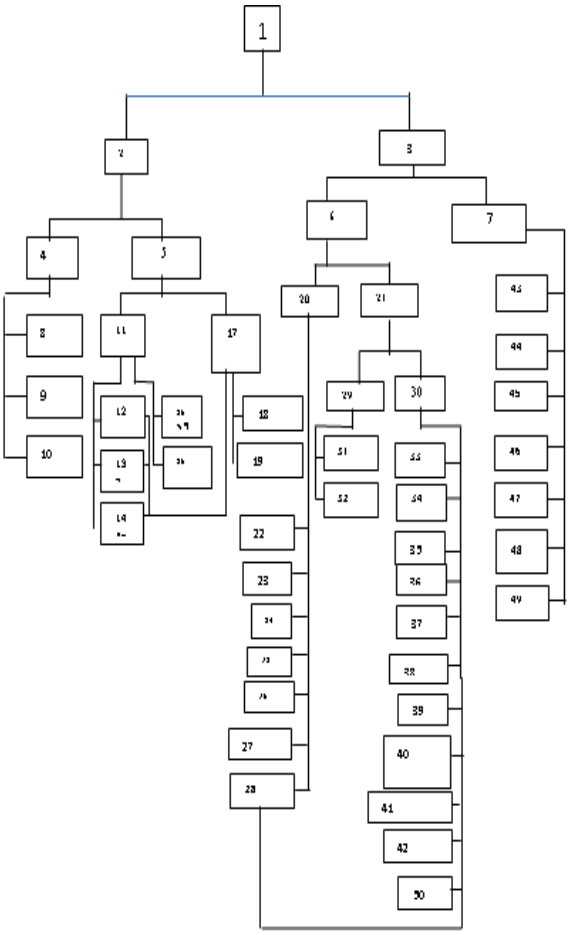


Figure 6. Example for Verb OntSet

The following trees are the Ontology trees of Noun and Verb entities.

1.2 Ontology Tree of Noun Entities



1- Noun, 2-Living, 3-Non-Living, 4-Human, 5-Non-human, 6-Concrete, 7-Abstract, 8-Profession, 9-Relation, 10-Character, 11-Plants, 12-Land, 13-Water, 14-Desert, 15-Edible, 16-Inedible, 17-Animals/Birds, 18-Vertebrae, 19-invertebrate, 20-Natural, 21-Man-Made, 22-Bodyparts, 23-Seasons,24-Places,25-Objects,26-Diseases, 27-Supernaturals, 28-Consumables, 29-Mobile, 30-Immobile, 31-Vehicles, 32-Machinery, 33-Building, 34-Jewellery, 35-Cutleries, 36-Furniture, 37-Weapons, 38-Instruments, 39-Clothes,40-Appliances & Gadgets, 41-Religious Materials, 42-Entertainment, 43-Emotion, 44-Time, 45-Event, 46-Act, 47-Festivals, 48-Senses, 49-Shapes, 50-Medicine

Figure 7. Ontology Tree of Nouns

1.3 Ontology Tree of Verb Entities

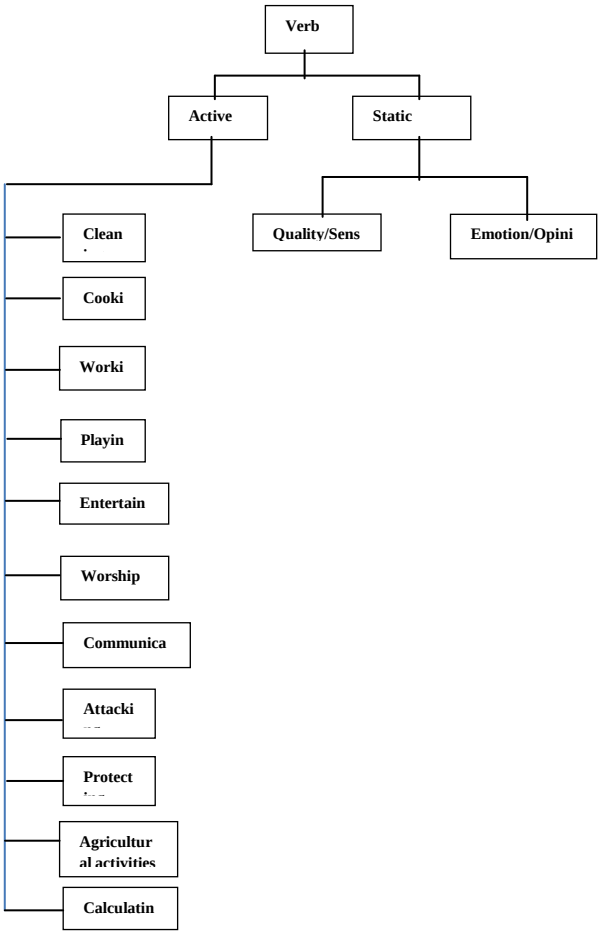


Figure 8. Ontology Tree of Verbs

1.4 Semantic Network between Noun and Verb Ontology Trees

Once the ontology trees are arrived at, the mapping of the categories within nouns and mapping of the categories between nouns and verbs are done based on their semantic relationships. This mapping results in a semantic network which facilitates the establishment of contextual meaning of a word in a given sentence.

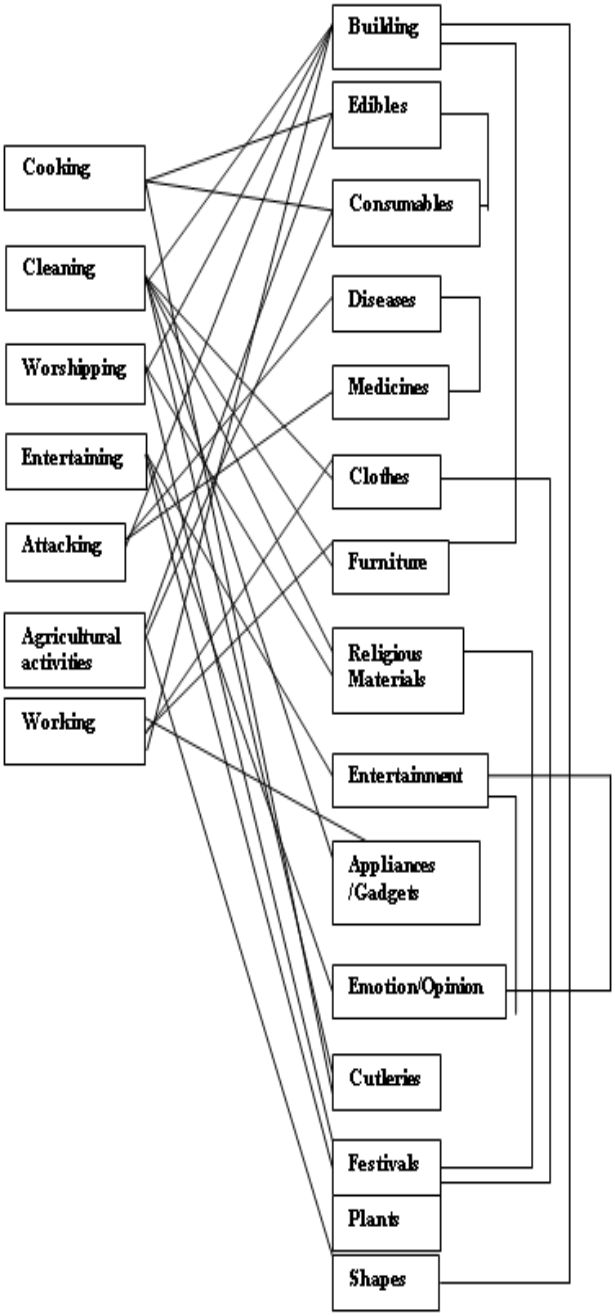


Figure 9. Establishment of a Semantic Network (OntNet)

1.5 Sub categorization features Tagging using OntNet

Using the above seen ontology trees of nouns and verbs and the OntNet derived, the constituent words are tagged with their semantic features as follows.

kiNNam 'Bowl – [Noun][+Non-living][+Concrete][+Man-made][+Immobile][+Cutleries]

Some nouns can have more than one set of semantic features.

koozi1 'Chicken – [Noun][+Living][+Non-human][+Animals/Birds][+Land]

This means the '*koozhi1*''Chicken is a living entity. It is non-human, a bird and it is a land inhabitant.

koozi2 'Chicken – [Noun][+Non-living][+Concrete][+Nature][+Consumables]

Tagging can be done for verbs too. The sub-categorization features of the verb ***kuththu*** 'punch give the characteristics of ***kuththu.***

kuththu 'Punch – [+Verb][+Active][+Attacking]

Certain verbs have more than one meaning. The verb ***yeNNu*** can be considered. It can be tagged with two different sets of semantic features based on its two possible meanings.

yeNNu1 'Think – [+Verb][+Static][+Emotion/Opinion]

yeNNu2 **'**Count - [+Verb][+Active][+Calculating]

Some words can fall under more than one category (Category ambiguity).

nei 'Ghee - [Noun][+Non-living][+Concrete][+Man-made][+Consumables]

nei 'Weave - [+Verb][+Active][+Working]

WORD SENSE DISAMBIGUATION USING ONTNET (ONTOLOGY AND WORDNET)

This section explains how OntNet is successfully used in Word Sense Disambiguation.

Word Sense Disambiguation

The following sentence can be considered.

avaN **[NP]** *veelai vaNnankinaaN* ‘ He worshipped the spear

The translation for the above sentence according to google is “He bowed down to work”. *veelai* has been translated as ‘Work’ which is wrong in this context. Now, using our OntNet for Tamil, the word sense of each of these words in the given context can be elicited and thereby arriving at the correct translation.

First, the sentence is tokenised. Then the constituent words are morphologically parsed. Each word is tagged with its corresponding POS tag and semantic features. Then syntactic parsing is done where VPs are located. After that, the semantic relationship between the noun and the verb in the VP is established. Based on this semantic relationship, translation is done.

After syntactic parsing, the output for the above seen sentence will be:

avaN **[NP]** *veelai vaNnankinaaN* **[VP]**

More than one set of sub categorization features are possible for *veelai* **and** *vaNnankinaaN***.**

veelai ‘ Work - [Noun][+Non-living][+Abstract][+Act]

veelai ‘ Spear -[Noun-Acc.case][+Non-living][+Concrete][+Manmade][+Immobile][+Religious]

vaNnankinaaN‘Worshipped(He) - [Verb-Finite][Active][+Worshipping]

Using the mapping of sub categorisation (Figure 9), it can be deduced that [+Act] is not related to [+Worshipping]. But [+Religious] is related to the category [+Worshipping]. Therefore the corresponding meaning is assigned to the word *veelai* as Spear. Thus OntNet plays a vital role in disambiguating the ambiguous words.

RESULTS AND CONCLUSION

The word level translator along with word sense disambiguator tool is tested with sentences from Tamil news papers and websites. Most wrong translations that occur in Google have been rectified using this tool. An approximate of 2000 nouns and 1000 verbs of Tamil language are semantically categorized in this tool. Almost 50 categories of noun and 20 categories of verbs are identified in this tool. A full-fledged knowledge base can be developed involving a larger number of OntSets. With minute categorization, it may be possible to resolve all the syntactic level ambiguities.

ACKNOWLEDGEMENT

I thank the University Grants Commission for funding this project under Minor Research Project scheme, 2014-2015.

REFERENCE

1. Andrew Carnie. Modern Syntax A Course book, 2011. Cambridge University Press.
2. Andrew Radford, English Syntax – An Introduction, 2004, Cambridge University Press, Cambridge
3. Anne Abeillé (ed), Treebanks – Building and using parsed corpora, Kluwer Academic Publishers
4. Bonnie Jean Dorr, Machine Translation : A view from the Lexicon, 1993, Publisher: The MIT Press, Cambridge, Massachusetts, London
5. Brian Roark and Richard Sproat. Computational Approaches to Morphology and Syntax, 2007. Oxford University Press.
6. Christopher D.Manning, HinrichSchutze, Foundations of Statistical Natural Language Processing, MIT Press, England

7. Daniel Jurafsky& James H.Martin, Speech and Language Processing, 2003, Pearson Education Inc.

8. Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.
9. Pon. Kothandaraman. A grammar of contemporary literary Tamil, 1997. International Institute of Tamil Studies.
10. Tyne Liang and Dian-Song Wu. Automatic Pronominal Anaphora Resolution in English Texts. Dept. of Computer and Information Science, National Chiao Tung University, Taiwan.
11. Akilandeswari, A and Sobha, Lalitha Devi, Resolution for Pronouns in Tamil Using CRF, AU-KBC Research Center, MIT Campus of Anna University. Chennai
12. Anaphora Resolution Using Named Entity and Ontology, Sobha, L, AU-KBC Research Center, MIT Campus of Anna University
13. Koehn, Philip and Knight, Kevin, Knowledge Sources for Word Level Translation Models, Infosciences Institute, University of Southern California.