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

Dr. R. Padmamala

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.

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.

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.
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**

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

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.

**METHODOLOGY**

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 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 *viiputhi* ‘Sacred ash’ is
grouped with the words *thiruniiRu* and *thuNNauRu*
which have the same meaning. These three words
together constitute a synset.

**Synset ID :** 9635 **POS :** NOUN

**Synonyms :** விபூதி, திருநீறு, துன்னூறு, துத்து

**Gloss :** ஒன்றை தறை மற்றும் புஜங்களின் மீது பூசிக்ககொள்ளும் ஏதொவது ஒரு ஹஊமத்திைிருந்து துைவி மூைமொக 

**Example statement**

"மகொத்மொஜி ஹநொ 

yொய்ப்பட்ட 

kுழந்றதயின் உடைில் விபூதி 

இடுகிைொர்"

**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)

**OntSet ID 1: NOUN**

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

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


1.3 Ontology Tree of Verb Entities

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.
words are tagged with their semantic features as follows.

\textit{kiNNam} ‘Bowl – [Noun]+Non-living][+Concrete][+Man-made][+Immobile][+Cutleries]

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


This means the ‘\textit{koozhi1}’ ‘Chicken is a living entity. It is non-human, a bird and it is a land inhabitant.

\textit{koozi2} ‘Chicken – [Noun]+Non-living][+Concrete][+Nature][+Consumables]

Tagging can be done for verbs too. The subcategorization features of the verb \textit{kuththu} ‘punch give the characteristics of \textit{kuththu}.

\textit{kuththu} ‘Punch – [+Verb][+Active][+Attacking]

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

\textit{yeNNu1} ‘Think – [+Verb][+Static][+Emotion/Opinion]

\textit{yeNNu2} ‘Count – [+Verb][+Active][+Calculating]

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

\textit{nei} ‘Ghee – [Noun]+Non-living][+Concrete][+Man-made][+Consumables]

\textit{nei} ‘Weave – [+Verb][+Active][+Working]

\textbf{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


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. Akilanadswari, 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.