LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 14:1 January 2014
ISSN 1930-2940

Resolution of Lexical Ambiguity in Tamil

S. Rajendran, Ph.D.

==

1. Introduction

Language is burdened with ambiguity; a single utterance can have a number of interpretations or meanings. The native speakers who speak a natural language have an implicit knowledge or competence to understand correctly these ambiguous utterances. They are capable of assigning an interpretation to any of the utterances they generate. They not only assign an interpretation to every utterance in their language, but also know that there are utterances that may have more than one semantic interpretation. These utterances are usually referred to as ambiguous utterances. When an utterance has more than one interpretation, it is usually referred to as ambiguous. Ambiguity means that utterances have same form but have different interpretations. Ambiguity may result from two homonyms/homographs occurring in the same structural position, as in the following example.

1. *avan kaal pakutiyaic caappiTTaan*
   ‘He ate quarter of something’/’He ate the leg part of something’

The sentence is ambiguous as the word *kaal* can mean ‘quarter of’ or ‘leg’. It may also occur when constituents in larger structures have more than one interpretation according to their internal structure and syntactic position.

2. *veLLai maruntu kuppi*
   ‘medicine bottle which is white in colour/a bottle with white medicine’

The sentence is ambiguous because the word *veLLai* ‘white’ can attribute either *maruntu* ‘medicine’ or *kuppi* ‘bottle’ The first one is called lexical ambiguity and the second structural ambiguity. Lexical ambiguity refers to the type of ambiguity those results from the occurrence of homonyms/homographs. Let us look at a few lexical ambiguity resolutions taking Tamil as the target language.

2 Lexical Ambiguities

The lexical ambiguity is a very common type of ambiguity. It includes, for example, the nouns such as *paTi* ‘step (of a stair)’/ ‘a kind of measure’, *kuTi* ‘drinking habit’/ ‘people’, *maTam* ‘foolishness’/ ‘mutt’ , etc, verbs such as *piTi* ‘catch’ / ‘to like’ , *kaTi* ‘bite’/ ‘to rebuke’ , *muTi* ‘to knot’/ ‘to finish’, *paTu* ‘to lie down’/ ‘to suffer’, etc and the adjectives such as *virinta ‘wide’/ ‘that which has blossomed’, kuRainta* ‘less’/ ‘that which has reduced’, *veLutta ‘white’/ ‘that which has become white’, kaRutta* ‘black’/ ‘that which has become black’, etc. There are tests for establishing lexical ambiguity. One of the tests is, for example, for the word *kaTinamaana* there are two opposite words, *metuvaana and eLitaana*. Consider the following example,

3a. kaTinamaana miTTaayaik kaTikka mutiyaatu
‘You cannot bite a hard sweet’

3b. kaTinamana collukkup poruL kuuRa iyalaatu
‘You cannot give meaning to a hard word’

The reason for this ambiguity is that the word has more than one meaning. But it is not clear when there is only one word involved in ambiguity. Though the noun *paTi* ‘a measure’ and the verb *paTi* ‘to study’ have same spelling/pronunciation they are two different words. They are examples of homophones/homographs. One may wonder whether the noun *kaTi* and the verb *kaTi* are examples for homonyms/homographs or not. Doubt may arise whether the word *mutal* in *mutal maaNavan* ‘first student’ and *aintu mutal* ‘from five’ are one and the same or not. To tell that one shows lexical ambiguity and the other homonymy/homography is not correct for all. This may be accidental.

There are three basic types in lexical ambiguity: category ambiguity, ambiguity due to homography and ambiguity due to ploysemy.

2.1 Category Ambiguity

Category ambiguity is the most straight forward type of lexical ambiguity. This happens when a given word is be assigned to more than one grammatical or syntactic category as per context. One can find a number of such examples in Tamil. For example the word *paccai* ‘green’ can be both noun as well as adjective. Similarly the word *cuTu* can be both a verb as well as an adjective. *kaTi* could be both a verb as well as a noun. The words like *meelee* and *kiizee* could be adverbs and postpositions.

4a. avan meelee cenRaan (adverb)
‘He went up’

4b. avan meecai meelee niRkiRaan. (postposition)
‘He is standing on the table’

Category ambiguities can often be resolved by morphological inflection. For example, *aTi* in *avan aTikkiRaan* ‘he is beating’ is a verb and *aTi* in *avanaal anta aTiyait taangka muTiyavillai* ‘He could not bear that beating’ is a noun. Frequently category ambiguity can be resolved by syntactic parsing. However, the problem increases when several categorically ambiguous words occur in the same sentence, each requiring being resolved syntactically.

2.2 Homography and Polysemy

If two entirely different words have different meanings the ambiguity arises due to homography. In the following example the word *paTi* shows homography.

5a. avan tantai avaniTam nanRaakap paTi eRu kuuRinaar
‘His father told him to study well’

5b. avan paTi vaziyaaka meelee eeRinaan.
‘He climbed up through steps’

Similarly *aTTai* can denote ‘leech’ as well as ‘binding’.

6a. avan puttakattin aTTaiyaik kizittu eRintaan
‘He has torn away the binding of the book’

6b. avan aTTaiyaik konRaan
‘He killed the leach’

If a word has two or more meanings it can be said that the ambiguity is due to polysemy. Polysemy expresses extension of meaning. The polysemous words may express new meanings by metaphoric or metonymic extensions. For example the word *kiLai* ‘branch’ may denote branch of a tree as well as a branch of a bank. *naTa* can denote the action of walking as well as happening or functioning of something.

7a. avan tinamum kaalaiyil paLLikku naTantu celkinRaan
‘He goes to school daily by walking’

7b. anta niRuvanam nanRaaka natantukoNTirukkinRatu
‘That organization is functioning well’

7c. anta tiyeeTTaril cinimaa naTantukoNTirukkinRatu
‘A cinema is running in the theatre’

ooTu can denote the human action of running as well flowing of a river.

8a. avan viraivaaka ooTukiRaan
‘He is running fast’

8b. tanjaavuur vaziyaaka kaaviriyaaRu ooTukiRatu
‘The river Kaviri flows through Thanjavur’

kaN may denote the eye of animate beings as well as the eye-like spot in the coconut.

9a. avan tan kaNkaLai muuTinaan
‘He closed his eyes’

9b. teengkaaykku muunRu kaNkaL uNTu
‘There are three eye-like spots in the coconut’

In the following sentence the word *keeL* denotes both the perception through ears as well as ‘asking’.

10a. raatai raajaa keeTTatai avaniTam kuuRinaaL
'Radha told him what Raja has asked her'
'Radha told him that Raja has heard that'

This sentence is ambiguous giving a number of interpretations; the following could be at least two interpretations.

10b. raatai raajaa tan kaataal keeTTatai avaniTam kuuRinaaL
'Radha told him what Raja has heard by his ears'

10c. Raatai raajaa vinaviyatai avaniTam kuuRinaaL
'Radha told him that Raja had asked her'

Sometimes among the homographs, the use of one may be greater than the other. In that case the ambiguity can be resolved on the basis of text. This is done by setting aside the unusual meaning form the dictionary unless it is required for translation.

As for as machine translation is concerned both the homography and ploysemy are treated alike, as the aim is to find out the meaning by context. The homographs belonging to different grammatical categories can be resolved as explained before. But if they belong to the same grammatical category syntactic parsing may not be enough. One common approach is to assign semantic features such as 'human', 'female', 'liquid' etc and to specify which features are compatible in the given syntactic constructions, via selection restrictions. For example it might be specified that the verb *kuTi* 'drink' has an 'animate' subject and a 'liquid' object.

2.2.1 Homography in inflected words

The homography can be resolved by different morphological analysis. The following examples will reveal this.

11a. avan kaTalai tinRu makizntaan
'He enjoyed eating pea nut'

11b. avan kaTalai kaNTu makizntaan
'He enjoyed seeing the sea'

In the first sentence the noun *kaTalai* denotes 'pea nut' and in the second case *kaTalai* has to be analysed as *kaTal* + *ai* (accusative case marker) and interpreted as *kaTal* 'sea'.

As in the case of the following example, the inflected word of one type of morphological analysis resembles an inflected word form of another morphological analysis, there by showing ambiguity due to homography.

12a. avan (tuNi) neytaan.
‘He weaved (cloth)’

12b. avan neytaan virumpukiRaan
‘He likes gee only’

In the first sentence the word *neytaan* has to be interpreted after analyzing it into *ney* ‘weave’+*t* (past tense)+*aan* (third person masculine singular) and in the second sentence *neytaan* has to be interpreted as *ney* ‘ghee’ + *taan* ‘only’. Even the two root words *ney* ‘weave’ and *ney* ‘ghee’ are homogrpahs showing lexical ambiguity.

2.2.2 Homography Due to Historical Functional Reorganization

The inflected forms of nouns or verbs will denote different word category or functional category due to historical meaning change. For example many of the postpositions in Tamil are historically the inflected forms of verbs. The inflected forms *iruntu* ‘from’ , *paRRi* ‘about’, *kuRittu* ‘about’, *oTTi* ‘about’, *koNTu* ‘by (means of)’, *vaittu* ‘by (means of)’, *cuRRi* ‘around’, *nookki* ‘towards’, *munti* ‘before’, *viTa* ‘than’ , and *kuuTa* ‘along with’ are the inflected forms of the verb *iru* ‘be’, *paRRu* ‘catch’, *kuRi* ‘aim’, *oTTu* ‘stick’, *koL* ‘have’, *vai* ‘keep’, *cuRRu* ‘go aroung’, *nookku* ‘look at’, *muntu* ‘over take’, *viTu* ‘leave’, and *kuuTu* ‘assemble’ respectively.

13a. avan viiTTil-iruntu veLiyeeRinaan (*iruntu* – postposition)
‘He went out from the house’

13b. avan viiTTil iruntu vantaan (iruntu – participle form of the verb *iru* ‘be)
‘He was in the house (habitually/continuously)’

14a. avan avaLaip paRRi peecinaan (*paRRi* – postposition)
‘He talked about her’

14b. avan avaL kaiyaip paRRi izuttaan (participle form of the verb *paRRu* ‘hold’)
‘He caught hold of her hand and pulled it’

15a. avan avaLaik kuRittup peecinaan. (*kuRittu* – postposition)
‘He talked about her’

15b. avan avaL colvataik kuRittu vantaan (*kuRittu* – participle form of the verb *kuRi* ‘note down’) ‘He was noting down what she was telling’

16a.avan anta talaippai oTTi peecinaan. (*oTTi* – postposition)
‘He talked about that title’

16b. avan poosTar oTTi pizaikkinRaan. (*oTTi* – participle form of the verb *oTTu* ‘stick’)
‘He ekes his livelihood by pasting posters’

17a. avan katti koNTu atai veTTinaan (*koNTu* - postposition)
‘He cut it with a knife’

17b. avan pencilaic ciivik-koNTu peecinaan. (*koNTu* – participle form of the verb *koL* ‘have’) ‘He was speaking while sharpening the pencil’

18a. avan katti vaittup pazam veTTinaan (*vaittu* - postposition)
‘He cut the fruit with a knife’

18b. avan paNam kaiyil vaittuk-koNTu cuutaaTinaan. (*vaittu* – participle form of the verb *vai* ‘keep’) ‘He gambled by keeping the money at hand’

19a. avan viiTTaic cuRRi marangkaL niRkinRana (*cuRRi* –postposition)
‘The trees are standing around his house’

19b. avan avaLaiyee cuRRi varukinRaan (*cuRRi* –participle form of the verb *cuRRu* ‘go around’) ‘He is going after her’

20a. avan avaLai nookki naTantaan (*nookki* –postposition)
‘He went towards her’

20b. avan avaL mukattai nookkic cirittaan (*nookki* – participle form of the verb *nookku* ‘look at’ ‘He smiled looking at her face’

21a. avan avaLukku munti angku vantaan. (*munti* – postposition)
‘He came there before her’

21b. avan avaLai munti naTantukoNTiruntaan. (*munti* –participle form of the verb *muntu* ‘overtake’)
‘He was walking overtaking her’

22a. avan avaLai viTa nallavan (*viTa* – postposition)
'He is better than her'

22b. avan avaLai viTa virumpavillai (*viTa* – infinitive form of the verb *viTu* 'leave')
'He does not want to leave her'

23a. avan avaL kuuTa vantaan. (*kuuTa* – postposition)
'He came with her'

23b. avan avarkaLuTan kuuTa virumpinaan (*kuuTa* - participle form of the verb *kuuTu* 'gather together'
'He wanted to gather together with them'

The word *enRu*, which is the inflected from the verb *en* 'to say', has two different grammatical functions thus showing ambiguity due to homography.

24a. avan nallavan enRu ninaitteen (*enRu* functions as complementizer)
'I thought that he is a good person'

24b. avan tiTiir enRu inku vantaan (*enRu* functions as adverbializer)
'He came here suddenly'

The word *enRu* 'when' is having homographic relation with the inflected verbal form *enRu*.

24c. avan enRu varukiRaan

'When does he come?"

The inflected verbal forms which can be analyzed as verb+*um* (future suffix) can be interpreted at least in two ways.

25a. atu naaLai varum
'It will come tomorrow'

25b. atu varum naaL enakkut teriyaatu
'I don't know the date of its coming'

The inflected verbal form which can be analyzed as verb + tense + *atu* can be interpreted in three ways.

26a. atu vantatu 'it came' (*vantatu* is the finite verbal form of the verb *vaa* 'come')

26b. atu vantatu enakkut teriyaatu (*vantatu* is the gerundival form)
“I did know that it had come’

26c. anta ceytittaaL neeRRu vantatu (*vantatu* is the participial noun form of the verb *vaa* ‘come’)
‘That newspaper is yesterday’s one’

The ambiguity can be resolved by selectional restriction, context, collocation, co-occurrence, etc.

3. Resolution of Lexical Ambiguity

Resolution of ambiguity is the central problem in language comprehension as well as natural language processing applications. As language speakers, we resolve the lexical ambiguity by looking at the context. The context need not be the immediate one. Even distance context or the topic of discourse can also help us to resolve ambiguity. The selection of apt sense is a challenging job as many rules are needed to select the appropriate sense by context or collocation or co-occurrence. Though the method of manipulating the context varies from linguistic analysis to automatic computational analysis, the concern is common for both, i.e. capturing context.

Resolving lexical ambiguity can involve different kinds of information. “These include word-specific information such as morphological information, part of speech (the syntactic category of the word), relative sense frequency (preferred sense, either generally or based on domain), semantic features (the semantic components, often drawn from a potentially large set of primitives, that contribute to meaning) as well as contextual information such as syntactic role (e.g., a particular sense may be the only one allowed as the object of a given preposition), role-related preferences (selectional restrictions defining relations between a noun and verb), semantic relations (most usually, senses of or associations with surrounding words), etc.” (Ide and Véronis, 1990) It has recently been suggested that an effective word sense disambiguation procedure will require information of most or all these types (McRoy, 1992). However, most methods utilize only one or two of the potential information sources listed above.

3.1 POS Tagging

Category ambiguity can be resolved by POS tagging. Ambiguities of syntactic category are resolved as part of the process of ‘parts-of-speech tagging’; parts-of-speech tagging involves labeling each word in input sentence with its category; it is the first stage of processing in many applications of natural language processing. The rules of grammar constrain the allowable sequences of syntactic category. Consequently, the category of a word can be resolved with a high degree of accuracy just by looking at the categories of a few preceding words. We have seen under categorical ambiguity that there are a number of words in Tamil which are homographic pairs denote different meanings as they belong to different POSs. The categorical ambiguity can be resolved by POS tagging. For example we have seen that the word form *ney* has two meanings depending on the category to which it belongs; as a noun *ney* means 'ghee' and as a verb *ney* means 'weave'. The ambiguity of the suffix *taan* due to homograpy (as noted earlier) can be resolved by POS labeling. *taan* attached to nominal *ney* 'ghee' gives emphatic meaning, whereas *taan* attached to the verbal *ney* gives the meaning 'PAST-he'.

Similarly the functional shift of the inflected forms of certain verbs listed above (*iruntu* 'from/having been' , *paRRi* 'about/having caught'', *kuRittu* 'about/having note down', *oTTi* 'about/having stuck', *koNTu* 'by (means of)/having', *vaittu* 'by (means of)/having kept', *cuRRi* 'around'having going around', *nookki* 'towards/having looking at', *munti* 'before/having gone before', *viTa* 'than/to leave', and *kuuTa* 'along with/to increase') can be resolved by POS-tagging. The two different functions of them as verb and postpositions can be resolved by POS tagging them respectively as verb or postposition. Similarly the two different meanings of the form *paTi* 'read (verb)', 'step (noun)' can be resolved by POS tagging. Use of *enRu* as pure verb, complementizer can be resolved by the same means.

3.2 Selectional Restriction

Selectional restriction is widely used for resolving lexical ambiguity (Katz and Fodor 1963). Selectional restrictions are the semantic constraints the word sense may impose on the sense of other words that combined with it. In other words, selectional restrictions are semantic requirements associated with the structures representing meanings of words or phrases, which must be met by another semantic structure before the two can be combined. For example, in Tamil the verb *tin* 'eat' in its literal sense requires its subject be an animate being and its object be some edible thing; so in the sentence *atai oru vilangku tinRatu* 'an animal/handcuff ate it', the word *vilanku* in this sentence is interpreted as 'animal' rather than 'handcuff'. In general, selectional restrictions are one-place predicates that test for the presence or absence of some semantic feature, or some Boolean function of such predicates.

The use of selectional restrictions in disambiguation is, in principle at least, quite straight forward. One simply has to select the sense (or senses) of a word that selectional restrictions will allow to combine with other semantic structures in the sentence; this is possible as it fulfills the requirements of those other structures, or because it fulfills its own requirements. There are difficulties in finding semantic features that can be used consistently and specifying the selection restriction for nouns and verbs based on these features. Even then these are widely used in machine translation system often in combination with case roles. But the semantic features cannot solve all the problems, even in situations for which they have been devised. For example, let us take the word *aTTai*. As we have indicated earlier that it is used in the senses of ‘binding’ and ‘leech’. These two senses can be differentiated explaining the relevant co-occurrence restrictions we find out in the following sentences in which *aTTai* is used.

27a. puttakattin aTTai kizintuviTTatu
‘The binding of the book is torn’

27b. aTTai uurntu celkinRatu
‘The leech is crawling’

The verbs like *kizi* ‘tear’ will take the objects like *aTTai* ‘binding’ which can be torn as their subjects and the verbs like *uurntucel* takes subjects like *aTTai* ‘leech’ which can crawl.

Similarly the two different senses of *nuul* ‘book/thread’, *vilangku* ‘animal/handcuff’, *maalai* ‘garland/evening’ can be resolved by selectional restrictions. Look at the following examples.

28a.avan anta nuulai vaacittu muTittaan
‘He finished reading that book’

28b. avan nuulai tuNiyaaka neytaan
‘He weaved the thread into cloth’

29a. avan avaL kaiyil vilangku maaTTinaan
‘He put the handcuff in her hand’

29b. avan viiTTil puunai naay poonRa vilangkukaLai vaLarkkiRaan
‘He is grooming the animals such as cat and dog’

30a. avan avaL kazuttil maalai iTTaan
‘He put a garland around her neck’

30b. avan maalaiyil viiTu tirunmpinaan
‘He returned home in the evening.

Selectional restriction will not be helpful to disambiguate words in the absence of sense selecting words. For example in the following example it is not possible to give the proper reading to *nuul* ‘book/thread’.

31.avan nuul vaangkinaan

‘He bought book/thread’

The use of selectional restriction for natural language system needs a knowledge base of selectional restrictions pertaining to each word sense. Such knowledge base does not exist for Tamil and it is difficult to build too. One such attempt has been made by the FrameNet project (Johnson and Fillmore 2002). Attempts have been made to make use of WordNet too for the same purpose.

3.3 Neighboring Words

There exists a relation between the ambiguous word and the neighbouring words in a text, that is, there exists a general semantic relationship between one of the candidate senses and nearby words in the text. Many methods of word sense disambiguation aims at capturing this cue. It is always context which decides the meaning of a word. It can be told flagrantly that a word cannot exist without context.

The dependency of meaning on context can be proved with a large amount of examples from Tamil. For example, let us take the word *maalai*; the proximity of the word *puu* ‘flower’ with *maalai* gives us the cue that it is ‘garland’ sense of *maalai* which is projected in this context and not the ‘evening’ sense of *maalai*. The topic of text or the domain of the text as a whole can be a helpful cue. The problem we face is using these clues precisely to determine the semantic relationship and there by select the correct sense. Context clustering approach based on the idea of word space or vector space (Schutze 1998) exploits cues from the neighboring words. This can be easily attempted for Tamil as it is a corpus dependent unsupervised method. The results are encouraging (Baskaran 2002, Rajendran and Anandkumar 2013).

3.4 Dictionary Definitions

The primary function of a dictionary is to provide the userr with the possible meanings or senses of a word. The dictionary makes use of definitions to fulfill its mission. The meanings of a word are explained by making use of already known words or simple words. This quality of a dictionary can be exploited for the resolution of lexical ambiguity. Lesk (1986) proposed to use the dictionary definitions to disambiguate the context. The definitions found in the machine readable dictionaries (MRDs) gives us contextual words which can be utilized for disambiguating the word senses. The context available in the dictionary definitions of words can be visualized as a bag of words. These words can be matched against the context in which the target word appears and there by select the correct sense from the candidate senses. The bag of contextual words need not be structured or in a proper word order pertaining to the target word. Lesk method offers us a simple method against many available complex methods. But the definitions given in the MRDs are not enough to disambiguate the word senses.

Let us look at the definition of *nuul* in the Tamil MRD *kriyaavin taRkaalat tamiz akaraati* (KTTA) (which means Dictionary of Contemporary Tamil).

nuul pe. panjcu, kampaLi mutaliyavaRRait tirittu tayaarikkappaTum izai ‘the yarn prepared from cotton and wool’

The definition gives contextual clues such as *panju* ‘cotton’, *kampaLi* ‘wool’, *tiri* ‘to yarn’, *tayaarikkappaTu* ‘prepare’, and *izai* ‘yarn’. But the definition cannot give clue to disambiguate *nuul* in the following sentence.

32. avan nuulaal caTTai taittaan
‘He stitched the shirt using thread’

The context furnished by the definition of *nuul* does not match with the contextual words in the given sentence.

3.5 Bayesian Classification Method

Bayesian classification method is a complex method. Bayesian classification method also makes use of the disambiguation cues from the neighboring words. It classifies the words according to the competing senses of the ambiguous word to which they can be associated with. For example the ‘handcuff’ sense of *vilangku* in Tamil is associated with *tiruTan* ‘thief’, *kuRRavaaLi* ‘criminal’, *ciRai* ‘jail’, *pooliis* ‘police’, etc. whereas the ‘animal’ sense of *vilangku* is associated with *kaaTu* ‘jungle’, *puli* ‘tiger’, etc. We can compute the probability of any given word occurring in the proximity of each sense by looking at a very large corpus of text in which each word is tagged with its correct sense, and counting the number of times that each sense occurs with various other words in its proximity. Then the probability of each sense can be computed in the context of neighbouring words; when disambiguation is necessary, the sense with the greatest probability can be chosen; this can be done even if those words do not all indicate the same sense. This approach presumes that all the words in the context are conditionally independent of one another; the probability of seeing one word in context is independent of seeing any other word in the same context. Clearly, this is not true in practice as the words of related meaning tend to cluster together. However, the method gives reliable results.

However, this approach requires sense-tagged corpora as its training data. This leads to the limitation of this approach. The sense-tagged corpora are not available for Tamil. New methods are adopted to circumvent this limitation. Yarowsky (1992) proposed that naïve Bayesian classification could be used if the goal is to determine the topic with which the ambiguous word is associated with instead of finding the fine-grinded sense of the word. For example, instead of having to determine separately the probability of ‘handcuff’ sense of vilangku associated with *tiruTan* ‘thief’, *kuRRavaaLi* ‘criminal’, *ciRai* ‘jail’, *pooliis* ‘police’, etc. and the ‘animal’ sense of *vilangku* associated with *kaaTu* ‘jungle’, *puli* ‘tiger’, etc we need to determine that any word related to animal indicates one sense of *vilangku* and any word related to handcuff indicates another. This method may be useful in many applications such as information retrieval. But while this method avoids the need for a sense-tagged corpus, it still requires supervised training as its learning phase that is based on some predefined knowledge source.

Yarowsky (1995) has also proposed a method by which decision list for disambiguation can be learned by unsupervised training. A decision list is an ordered sequence of very specific conditions for classifying a word by meaning; for example a decision list for word *nuul*, might include the conditions ‘if the next word is *tuNi* the topic *tai* ‘sew’, if the next word is *nuulakam* ‘library’ the topic is *nuul* ‘book’. The list can be derived from an extremely large corpus; we can get an extremely strong cue or seed for the ambiguous word. Yarowsky’s method requires separate training for each ambiguous word, so in practice only a few words can be taken care of. The use of this method for all ambiguous words remains a daunting one.

3.6 Lexical Cohesion

Lexicon cohesion elaborated by Moris and Hirst (1991) can be made use of to resolve certain type of lexical ambiguity. The continuity of lexical meanings of words, which results in chains of related words, contributes to lexical cohesion. Lexical cohesion is the cohesion that arises from semantic relationships between the words. All that is required is that there are some recognizable relations between the words.

The *thesaurus of modern Tamil* (Rajendran 2000) provides a fine-grained database for identifying lexical cohesion between words or concepts. *Tamil wordNet* prepared under the DIT funded project entitled “Development of Dravidain WordNet: An Integrated WrodNet for Telugu, Tamil, Kannada, and Malayalam” offers a reliable database for lexical cohesion.

We can adopt for Tamil a classification of lexical cohesion provided by Halliday and Hasan (1976) based on the type of dependency relationship that exists between words. According to them there are two classes of relationship: class of reiteration and class of collocation. The identity of reference or repetition of the same word as well as the use of superordinates, subordinates, and synonyms manifest the class of reiteration. The semantic relationships between words that often co-occur manifest the class of collocation. The systematic semantic and the nonsystematic semantic relationship can divided them further into two categories of relationship.

Systematic semantic relationships can be classified into different types of relation which manifest as antonyms, members of an ordered set such as *{onRu* 'one'*, iraNTu* 'two'*, muunRu* 'three'*}*, members of an unordered set such as *{veLLai* 'white'*, kaRuppu* 'black'*, civappu* 'red'*}*, and part-to-whole relationships like *{kaNkaL* 'eyes'*, vaay* 'mouth'*, mukam* 'face'*}*.

The word relationship collocation like *{tooTTam* 'garden'*, tooNTu* 'digging'*}* is nonsystematic. From a knowledge representation point of view this type of relationship is the most problematic one. Such collocation relationships exist between words that tend to occur in similar lexical environments. Words tend to occur in similar lexical environments because they describe things that tend to occur in similar situations or contexts in the world. Hence, context-specific examples such as *{tapaal aluvalakam* 'post office'*, ceevai* 'service', *tapaal villai* 'stamps'*, koTu* 'pay'*, viTu* 'leave'*}* are included in the class. (This example is lexical cohesion specific to the context of service encounters.)

Another example of this type is *{kaar* 'car', *viLakkukaL* 'lights'*, tiruppam 'turning'}*. These words are related in the situation of driving a car, but taken out of that situation, they are not related in a systematic way. Also contained in the class of collocation are *word associations.* They include examples such as *{puujaari* 'priest'*, kooyil* 'temple'*}, {kuTikaL* 'citizen', *intiyaa* 'India'*}*, and *{ciiTTikai* 'whistle'*, niRuttu* 'stop'*}*. Again, the exact relationship between these words can be hard to classify, but there does exist a recognizable relationship.

Moris and Hirst (1991) lists two major reasons for the importance of lexical cohesion for computational text understanding systems: "1. Lexical chains provide an easy-to-determine context to aid in the resolution of ambiguity and in the narrowing to a specific meaning of a word. 2. Lexical chains provide a clue for the determination of coherence and discourse structure, and hence the larger meaning of the text."

Word meanings do not exist in isolation. Each word must be interpreted in its context. For example, in the context {*jin* 'gin', *aalkakaal* 'alcohol', *paanangkaL* 'drinks'}, the meaning of the noun drinks is narrowed down to alcoholic drinks. In the context {*muTi* 'hair', *curuL* 'curl', *ciippu* 'comb', *alai* 'wave'}, *alai* 'wave' means a hair wave, not a water wave or a physics wave. In these examples, lexical chains can be used as a contextual aid to interpret word meanings.

Often, lexical cohesion occurs not simply between pairs of words but over a succession of a number of nearby related words spanning a topical unit of the text. These sequences of related words will be called lexical chains. There is a distance relation between each word in the chain, and the words co-occur within a given span. Lexical chains do not stop at sentence boundaries. They can connect a pair of adjacent words or range over an entire text.

Cruse’s (1989) taxonomies, meronomies, hierarchies and non-branching hierarchies are worth considered in building lexical cohesion. His non-branching hierarchies which include chains (e.g. *tooL* ‘shoulder’, *meeRkai* ‘upper arm’ , *muzankai* ‘elbow’, *munkai* ‘forearm’, *maNikkaTTu* ‘wrist’ and *kai* ‘hand’ and helices (e.g. *njaayiRu* ‘Sunday’, *tingkaL* ‘Monday’, *cevvaay* ‘Tuesday’ *putan* ‘Wednesday’, *viyaazan* ‘Thursday’, *veLLi* ‘Friday’, and *cani* ‘Saturday’) too help us in lexical cohesion. They show linear ordering, cyclic ordering and serial ordering. The week days, names of seasons (*vacantam kaalam* ‘spring’, *kooTai kaalam* ‘summer’, *ilaiyutir kaalam* ‘autumn’, and *kuLir kaalam* ‘winter’) , colour terms (*civapppu* ‘red’, *uutaa* ‘purple’, *niilam* ‘blue, *paccai* ‘green’, *manjcaL* ‘yellow’ and *aaranjcu* ‘orange’) make a cycle of relations one following the other denoting different kind of lexical cohesion. The numerals, both cardinal and ordinal, show serial ordering. The different kinds of lexical relations explained by Cruse (1989) can be viewed as different kinds or types of lexical cohesion.

The following table shows the different types of lexical cohesion possible for nouns from the point of view of wordNet relations (lexical and semantic relations).

Relations	**Subtypes**	**Example**
Synonymy		*puttakam* ‘book’ to *nduul* ‘book’
Hypernymy-Hyponymy		*vilangku* ‘animal’ to *paaluuTTi* ‘mammal’
Hyponymy-Hypernymy		*pacu* ‘cow’ to *paaluuTTi* ‘mammal’
Holonymy-Meronymy	Wholes to parts	*meecai* ‘table’ to *kaal* ‘leg’
,,	Groups to members	*tuRai* ‘department’ to *peeraaciriyar* ‘professor’
Meronymy-Holonymy	Parts to wholes	*cakkaram* ‘wheel’ to *vaNTi* ‘cart’
,,	Members to groups	*paTaittlaivar* ‘captain’ to *paTai* ‘army’
Binary Opposites	Antonymic (gradable)	*ndallavan* ‘good person’ to *keTTavan* ‘bad person’
,,	Complementary	*pakal* ‘day’ *to iruavu* ‘night’
,,	Privative (opposing	*ahRiNai* ‘irrational’ to *uyartiNai* ‘rational’

	features)	
,,	Equipollent (positive features)	*aaN* ‘male’ to *peN* ‘female’
,,	Reciprocal Social roles	*vaittiyar* ‘doctor’ to *ndooyaaLi* ‘patient’
,,	Kinship Relations	*ammaa* ‘mother’ to *makaL* ‘daughter’
,,	Temporal Relations	*munnar* ‘before’ to *pinnar* ‘after’
,,	Orthogonal or perpendicular	*vaTakku* ‘north’ to *kizakku* ‘east’ and *meeRku* ‘west’
,,	Antipodal Opposition	*vaTakku* ‘north’ to *teRku* ‘south’
Multiple opposites	Serial	*onRu* ‘one’, *iraNTu* ‘two’, *muunRu* ‘three’, *ndaanku* ‘four’
,,	Cycle	*njaayiRu* ‘Sunday’ to *tingkaL* ‘Monday’ .. to *cani* ‘Saturday’
Compatibility		*ndaay* ‘dog’ to *cellappiraaNi* ‘pet’

The following table shows the different types of lexical cohesion possible for verbs from the point of view of wordNet relations (lexical and semantic relations).

Relations	**Definition/sub types**	**Example**
Synonymy	Replaceable events	*tuungku* ‘sleep’ → *uRangku* ‘sleep’
Meronymy- Hypernymy	From events to superordinate events	*paRa* 'fly' → *pirayaaNi* 'travel'
Troponymy	From events to their subtypes	*ndaTa* → *ndoNTu* 'limp'
Entailment	From events to the events they entail	*kuRaTTaiviTu* 'snore' *muyal* ‘try’ *tuungku* 'sleep'
“	From event to its cause	*uyar* ‘rise’ → uyarttu ‘raise’
“	From event to its presupposed event	*vel* ‘succeed’ → *muyal* ‘try’

"	From even to implied event	*kol* 'murder' → *iRa* 'die'
Antonym	Opposites	*kuuTu* 'increase' → *kuRai* 'decrease'
"	Conversensess	*vil* 'sell' → *vaangku* 'buy'
"	Directional opposites	*puRappaTu* 'start' → *vandtuceer* 'reach'

Hirst (1987) used a system called "Polaroid Words" to execute intrasentential lexical disambiguation in his earlier work. Polaroid Words makes use of a number of cues for lexical disambiguation. This includes syntax, selectional restrictions, case frames, and a notion of semantic distance or relatedness to other words in the sentences; a sense that holds such a relationship is favored over one that does not hold this relationship. Relationships are determined by marker passing along the arcs in a knowledge base. This approach is based on the intuition that semantically related concepts will be physically close in the knowledge base. So this can be achieved by traversing the arcs for a limited distance. But Polaroid Words consider the possible relatedness between words in the same sentence; trying to find connections with all the words in preceding sentences is too complicated and too likely to be led astray. This weakness in Polaroid Words is taken into account in lexical chains; lexical chains provide a constrained easy-to-determine representation of context for consideration of semantic distance.

2.7 Neural Network

Ide and Véronis (1990) explain in detail the use of very large neural networks for word sense disambiguation. Everyday dictionaries represent ready-made, highly connected networks of words and concepts. For example, in KTTA, the definition of *nuul* 'book' contains words such as *paTi* 'read', *aTTai* 'binding', *acciTu* 'print', *taaL* 'paper', *tokuppu* 'volume'. The definition of *paTi* 'read' contains words such as *ezutappaTu* 'be written', *vaarttai* 'word', *uccari* 'pronounce', and *arttam koL* 'understand'. The definition of *taaL* 'paper' contains *ezutu* 'write' and *accaTi* 'print', and so on. All of these connections obviously form a dense cluster of semantically related words.

The fundamental assumption underlying the semantic knowledge represented in these networks is that there are significant semantic relations between a word and the words used to define it. The connections in the network reflect these relations. There is no indication within the network of the nature of the relationships, although the presence of words with important and relatively fixed semantic relations to their headwords in dictionary definitions is well-known, and much work has been applied to identifying and extracting this information.

We can build large networks of words and concepts for Tamil as several dictionaries are available in machine readable form for Tamil (including those available in online). We can exploit the existing structure of dictionaries, in which each word is connected to one or more senses (roughly equivalent to concepts), and each sense is in turn connected to the words in its definition. If the words *nuul* 'book' and *nuul* 'thread' are fed to such networks containing all the connections in the KTTA, we can expect that the appropriate senses of both *nuul* 'book' and *nuul* 'thread' will be triggered because of the activation they receive through their mutual, direct connections to the word *ezutu* 'write' and *tai* 'stitch' respectively, as well as numerous other indirect paths. The book sense of *nuul* will be activated by words such as *paTi* 'read', *aTTai* 'binding', *acciTu* 'print', *taaL* 'paper', *tokuppu* 'volume' whereas the thread sense of *nuul* is activated by the words such as *kai* 'hand' *iyantiram* 'machine', *tiri* 'to yarn', *melliya* 'tiny' and *izai* 'yarn'. The sheer density of the connections between the two senses of *nuul* 'book' and *nuul* 'thread' will override any other spurious connections between these two words.

4 Conclusion

There are innumerable approaches to the resolution of lexical ambiguity. Almost all are tested for English. As English has rich source of lexical knowledge stored in proper format as databases, it is possible to attempt the resolution of lexical ambiguity by making use of various methods. Tamil which lacks the resourceful lexical databases suffers in this effort. It is still possible to find out avenues to resort to all these approaches for Tamil taking into account the limited databases getting built now-a-days. All the approaches aiming at lexical disambiguation captures the context by some means. Before attempting any method it is better to understand the intricacies involved in executing all these approaches and to build reliable databases to implement them.

===

References

Baskaran S. 2002. Semantic analyser for word sense disambiguation. MS thesis. Madras Institute of Technology, Anna University, Chennai 2002.

Baskaran S and Vaidehi V. 2003. Collocation based Word Sense Disambiguation using Clustering for Tamil. International Journal of Dravidian Linguistics 33(1): 13-28, Thiruvananthapuram, India.

Christopher R., Johnson C.R. and Fillmore C.J. 2000. The FrameNet tagset for frame-semantic and syntactic coding of predicate argument structure. In: Proceedings, 1st Meeting of the North American Chapter of Association for Computational Linguistics, Seattle, pp 56-62.

Cottrell G.W. 1989. A Connectionist Approach to Word Sense Disambiguation. Pitmann, London, UK.

Cottrell G.W. and Small S. L. 1983. A connectionist scheme for modelling word sense disambiguation. Cognition and Brain Theory, 6, 89-120.

Cruse D.A. 1986. Lexical Semantics. Cambridge University Press, Cambridge.

Dravidian WordNet: An Integrated Wordnet for Telugu, Tamil, Kannada and Malayalam. DIT funded on-going project.

Fellbaum C. (ed.). 1998. WordNet: An Electronic Lexical Database. MIT press, Cambridge, MA.

Gale W.A. Church K.W., and Yarowsky D. 1992. Using bilingual materials to develop word sense disambiguation methods. In: *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation.*

Gale W.A., Church K.W., Yarowsky D. 1993. A method for disambiguating word senses in a large corpus. Computers and the Humanities, Special issue on Common Methodologies in Computational Linguistics and Humanities Computing, Ide N. and Walker D., eds.

Halliday M.A. and Hassan R. (eds). 1976. Cohesion in English. Longman Group Ltd, London, U.K.

Hirst G. 1987. Semantic Interpretation and Resolution of ambiguity. Cambridge University press, Cambridge, UK.

Ide, N and Véronis J. 1990. Very large neural networks for word sense disambiguation. Proceedings of the 9th European Conference on Artificial Intelligence, ECAI'90, Stockholm , 366-368.

Katz J. and Fodor J. 1963. The structure of a semantic theory. *Language,* vol. 39, pp. 170-210.

Lesk M.E. 1986. Automatic sense disambiguation using machine-readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings, 5th International Conference on system Documentation, Toronto, pp. 24-26. Association for Computing Machinery, New York, NY.

McRoy S.W. (1992). Using Multiple knowledge Sources for Word Sense Discrimination. Computational Linguistics, 18, 1, 1-30.

Miller G. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235–312.

Morris J. and Hirst G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17, 1, 21-48.

Pustejovsky J. 1995. *The Generative Lexicon*. Cambridge, MIT Press, MA

Rajendran S. 2001. *taRkaalat tamizc coRkaLanjciyam* (Thesaurus for modern Tamil). Tamil University, Thanjavur.

Rajendran S., Anandkumar M. and Soman K.P. 2013. "Computational Approach to Word Sense Disambiguation in Tamil. Conference Papers, 12th International Tamil Internet Conference 2013, 35-46.

Schutze H. 1998. Automatic word sense discrimination. Computational linguistics 24, 1, 97-124.
Yarowsky D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the14th lnternational Conference on Computational Linguistics (COLING-92)*, pages 454- 460.

Dr. S. Rajendran
Professor of Linguistics
Department of Computational Engineering and Networking
Amrita Vishwa Vidyapeetham
Ettimadai
Coimbatore 641105
Tamilnadu
India
raj_ushush@yahoo.com