Certain issues in the Development of Telugu - Tamil Machine Translation : A view from the lexicon

[1]*Parameswari K*, [2]*Uma Maheshwar Rao G*, [3]*Krupanandam N*, [4]*Lavanya J*, [5]*Christopher M*
Center for Applied Linguistics and Translation Studies
University of Hyderabad, Hyderabad – 500046.
[1]*parameshkrishnaa@gmail.com*, [2]*guraohyd@gmail.com*, [3]*nityakrupa@yahoo.co.in*,
[4]*jlavanyacse08@gmail.com*, [5]*efthachris@gmail.com*

Abstract: Machine Translation (MT) is one of the interesting and challenging tasks of Natural Language Processing. In any Machine Translation system, understanding the pair of languages involved are vital. The present work focuses on certain issues in the development of Telugu-Tamil Machine Translation from the point of the languages involved and the dictionaries that are used in Telugu-Tamil Machine Translation System which are unique since they are based on concept. The paper deals with the compilation of concept based dictionary for Machine Translation purpose and also deals with the divergences arise due to the differences in the lexemes of Telugu and Tamil.

1. Introduction:

Tamil, the South Dravidian Language and Telugu, the South Central Dravidian language are major languages of South India. The Machine Translation between Telugu-Tamil is a best example case taken for the development of MT since there is a great demand for the Translation of texts of each of these languages. Normally, Machine Translation is a challenging task where computers take over the task of translating one language into another. Though the languages involved viz. Telugu – Tamil are closely related, exhibit a number of dissimilarities in their linguistic behavior thus making the task a non trivial one.

The paper deals with the issues in the development of an automatic Telugu-Tamil Machine Translation System which is being developed under the project of IL-IL MT at CALTS, University of Hyderabad as part of the Consortium of Indian Languages to Indian languages Machine Translation Systems funded by DIT, Ministry of Information Technology, Government of India.

The lexical resources are essential for building any Machine Translation system. The Lexicon used in the building of Telugu-Tamil MT is one of the Machine Readable Dictionary types, which differs from printed conventional dictionaries of everyday use. The conventional dictionary is usually meant for defining and providing description about a lexeme. However, the concept based dictionary which is currently used contains lexemes without any encyclopedic knowledge.

2. Concept based Dictionary or Synset:

A concept is an idea which is language specific and based on the ontology of lexemes in languages. The concept Based dictionary is a component of a multilingual dictionary developed for 11 languages: *English, Hindi, Bengali, Marathi, Punjabi, Urdu, Tamil, Kannada, Telugu, Malayalam* and Oriya (Cf. Mohanty et.al.) by different Indian Institutions and are used in NLP applications. The greatest advantage of using synset is the conceptually related words are grouped under a single concept and the equivalents in the target language along with linkages provided. Here, Hindi is used as a pivot language and other synsets in other languages are built on the principle of translational equivalence. The Telugu-Tamil Machine Translation System uses Telugu and Tamil synsets which are developed by CALTS' NLP group and AU-KBC NLP group respectively. A lexeme used to express a concept in a language may not have the same meaning in all the contexts. The same lexeme may be found in different contexts expressing different meanings or concepts. For instance, a lexical item in Telugu corresponds to one or more lexical (sense's) items in Tamil.

In Telugu, the word *kuttu*^{1} is translated in Tamil as,

a. *kati* in the context of *cIma kuttu* 'to bite as an ant',
b. *wE* in the context of *battalu kuttu* 'to stitch clothes' ,
c. *kuwwu* in the context of *ceVvulu kuttu 't*o pierce ears'.

Here the question of providing an appropriate equivalent for *'kuttu'* requires word sense disambiguation. The concept which is the central point of this lexeme can help to avoid this problem. The dictionary which we are proposing as a suitable one for Machine Translation is of concept centered one rather than of one to one lexical matching. A word X in a language is taken as a concept, and the conceptually related words of X are provided as W1,W2,...Wn. The hierarchy of frequency is followed in an ascending order of giving equivalents. The links (L) are created between the source language and the target language lexemes. The concept Dictionary is used to perform a lexical transfer of the following:

(a) Situation (1) One to One :
Here a single lexical item is linked with a corresponding lexical item in Tamil.

Ex: X (sw1/L1 <--> tw1)

(where **X** is a context with category, **sw** is a source word, **L** is link, **tw** is target word)
Ex:
Telugu :
ID :: 7350
CAT :: NOUN
CONCEPT :: ఎవరికైనా అప్పు ఇచ్చినప్పుడు లేదా బ్యాంకు మొదలైన వాటిలో కూడబెట్టిన డబ్బుకి బదులుగా ఆ సమయం వరకు ఇచ్చే నిశ్చిత ధనము
EXAMPLE :: "శ్యాం వడ్డీకి డబ్బులు ఇస్తాడు"
SYNSET-TELUGU :: వడ్డీ /TAM1
This is the link to Tamil.

Tamil:
ID :: 7350
CAT :: NOUN
CONCEPT :: வட்டி
EXAMPLE :: "வங்கியில் வாங்கிய கடன் தொகைக்கான வட்டி குறைந்துள்ளது."
SYNSET-TAMIL :: வட்டி

(b) Situation (2) Many to One :

Here multiple lexemes are displayed with linkages with a single lexical item

Ex: X(sw1/L1, sw2/L1, sw3/L1, sw4/L1, sw5/L1 <--> tw1)

Telugu :
ID :: 73
CAT :: NOUN
CONCEPT :: అంతర్గతంగా కలిగి ఉన్న క్రియ
EXAMPLE :: "అందంలో అందంగా ఉండే భావం ఉన్నది"
SYNSET-TELUGU :: భావం/TAM1, భావార్థం/TAM1, భావన/TAM1, అర్థం/TAM1, తాత్పర్యం/TAM1

Tamil:
ID :: 73
CAT :: NOUN
CONCEPT :: இப்படிப்பட்டது அல்லது இப்படிப்பட்டவர் என்பதை அறிவதற்கான அம்சம்

EXAMPLE :: "மனிதனிடம் மனித தன்மை காணப்படும்."
SYNSET-TAMIL :: தன்மை

(c) Situation (3) One to Many :

Here a single lexeme is linked with multiple lexical items.

Ex : X(sw1/L1 <--> tw1, tw2,tw3)

Telugu :
ID :: 12019
CAT :: NOUN
CONCEPT :: ఒక వస్తువుని దగ్గరికి లాగే స్థితి
EXAMPLE :: "అయస్కాంతాలకు ఆకర్షణ శక్తి ఉంటుంది/తన కళ్లల్లో ఆకర్షణ ఉంది”
SYNSET-TELUGU :: ఆకర్షణ/TAM1

Tamil:
ID :: 12019
CAT :: NOUN
CONCEPT ::
EXAMPLE ::
SYNSET-TAMIL :: கவர்ச்சி, வசீகரம்,, ஈர்ப்பு

(d) Situation (4) Many to Many :

Here many lexical items are linked with many in the target side.

Ex: **X(sw1/L1, sw2/L2, sw3/L3, sw4/L4, sw5/L5 ,sw6/L6, sw7/L7 <--> tw1,tw2,tw3,tw4,tw5,tw6)**

Telugu:
ID :: 12833
CAT :: VERB
CONCEPT ::
EXAMPLE :: ""
SYNSET-TELUGU :: ప్రారంభించు/TAM1, మొదలుపెట్టు/TAM2, లేవదియ్యి/TAM3, ఆరంభించు/TAM4, ప్రారంభంచెయ్యి/TAM5, ఆరంభించు/TAM6, లేవదీయు/TAM7

Tamil :
ID :: 12833
CAT :: VERB
CONCEPT :: எழுப்பு, ஆரம்பி
EXAMPLE :: "நடு நடுவே அவன் ரமாவின் திருமணத்தைப் பற்றிய பேச்சை எழுப்பினான்"
SYNSET-TAMIL :: ஆரம்பி, தொடங்கு, எழுப்பு, ஆரம்பி, ஆரம்பம்_செய், துவங்கு, எழுப்பு

The dictionary uses the categories like Nouns, Verbs, Adjectives, Adverbs, Pronouns, Numerals, NST and Indeclinables such as Classifiers, Quantifiers, Interjections, Quotatives, Particles and Conjunctions. Other than this, a whole list of functional words like Case Markers and Tense, Aspectual and Model markers are also included in the dictionary. A bilingual dictionary, which is also a concept based one is used as a stand-by along with the synset dictionary in case of failing.

3. Divergences of Telugu and Tamil from the point of lexicon:

A translation divergence may occur when the underlying concept or “gist” of a sentence is distributed over different words for different languages. According to Dorr (1990), divergences are cross-linguistic distinctions in which the natural translation of one language into another results in a very different that of the original. She proposes seven types of divergent categories comprising of Thematic, Promotional, Demotional, Structural, Conflational, Categorial and Lexical Divergences. This classification of Dorr

(1990) on Machine Translation Divergence is taken as a base and is tried to map it with the Telugu-Tamil Machine Translation System. The paper focuses on three divergences due to lexical aspects of the languages involved in translation.

3.1 Conflational Divergence :

It occurs when the sense conveyed by a single word is expressed by two or more words in one of the languages. For instance, Telugu uses 'snAnaM ceVyyi' for 'bath' whereas it is expressed by 'kulYi' in Tamil.

II.a. TEL: nenu snAnaM ceswAnu.
'I bathing do-FUT-1p.sg'
TAM: nAnY kulYippenY.
'I bath-FUT-1p.sg'
ENG: I will bathe.

The Conflational Divergence is mainly carried out by the Multi Word Expression Module. Here the collocative words are given equivalents in the respective language. Multi-word expressions are a set of collocations of words which are often come with a non compositional semantics which otherwise could not be resolved. These forms are sequences of two or more words generally express a co-occurrence meaning. Telugu-Tamil Multi-Word Expression Module is built up with the database which consummates the words of co-occurance. The root form of the two or more sequences of words are used in the database. For instance, the following expression of Noun (N) and Verb (V) is carried out during the Telugu-Tamil Translation:

1.N N --> N N
uwwara praxeS, *uwwirap pirawecam*
Since 'uwwaraM' in isolation may mean either 'the north' or 'a letter', but in the context of the word indicating the name of a State, it needs to be listed.
2. N V --> N V
veru ceVyyi, pirivinYE ceVy
Here the word 'veru' may mean 'separation' and 'root'. But when it is followed by a verb like 'ceVyyi' it means 'to separate'.
3. N N N --> N N N
calana ciwra pariSrama,wirEp patac cafkam
Here the cinema is expressed in Telugu as 'calana ciwraM' i.e, 'motion picture' whereas in Tamil it is 'screen picture'.
4. N N --> 0 N
sahajaM sixXaM,0 iyarYkE
The term 'natural product' is expressed by two words in Telugu whereas in Tamil it is one.
5. N V --> 0 V
vidixi ceVyyi,0 wafku
xAdi ceVyyi, 0 wAkku

In Telugu, the intensifier compounds involve two words to express a single intensifyied form of the concept denoted by the nouns of temporal/spatial category whereas in Tamil by the corresponding reduplication of the head noun.
6. REDUP NST---> REDUP NST
moVtta moVxata, muwanY muwal

As described above lexemes of special cases such as phrases, idioms which are multi word expressions are taken care by the MWE module before the processing enters into the lexical transfer module.

Input :

```
<Sentence id="1">
1	((	NP	<fs af='samuxraM,n,,sg,,o,ti,ti' head="samuxra" vpos="vib1" name=1>
1.1	samuxra	NN	<fs af='samuxraM,n,,sg,,o,ti,ti' name="samuxra" ENAMEX TYPE="LOCATION"
SUBTYPE_1="LANDSCAPES">
	))
2	((	NP	<fs af='wIraM,n,,sg,,d,0,0' head="wIraM" drel=k7p:4 name=2>
2.1	wIraM	NN	<fs af='wIraM,n,,sg,,d,0,0' name="wIraM">
	))
</Sentence>
```

Output :

```
<Sentence id="1">
1	((	NP	<fs af='samuxraM,n,,sg,,o,ti,ti' head="samuxra" vpos="vib1" name=1>
1.1	0	NN	<fs af='^@0,n,,sg,,o,ti,ti' ENAMEX/TYPE="LOCATION" name="samuxra"
SUBTYPE_1="LANDSCAPES">
	))
2	((	NP	<fs af='wIraM,n,,sg,,d,0,0' head="wIraM" drel=k7p:4 name=2>
2.1	katarYkarE	NN	<fs af='^@katarYkarE,n,,sg,,d,0,0' name="wIraM">
	))
</Sentence>
```

3.2 Categorical Divergence :

Changes in category creates categorical divergence. It is due to the mismatch between the Parts of Speech Categories of the words involved in the pair of Languages. For instance, the word cAla in Telugu is ambiguously used as an adjective as well as an intensifier. However, in Tamil two distinct categories of words are used as shown below:

III.a. TEL: cAlA puswakAlu unYnYAyi.
'a lot book-pl being-3.p.pl.n'
TAM: nirYEya puwwakafkalY ulYlYanYa.
'a lot book-pl being-3.p.pl.n'
ENG: A lot of books are there.

III.b. TEL: cAlA eVwwugA uMxi.
'very high being'
TAM: mika uyaramAka ulYlYawu.
'very high being'
ENG: It is very high.

Handling Strategy :

(i) Repair Role :

{[cAlA<$cat>][N1.gA]} =>{[cAlA<INT>][N1.gA]}
{[cAlA<$cat>][N1]} => {[cAlA<ADJ>][N1]}

3.3 Lexical Divergence:

It arises when there is a lack of exact lexical equivalent but structure presents a translational equivalence between the language pair. Here, the literal translation of the source language word is substituted by a corresponding translational equivalent to resolve the problem.

For instance, IV.a. TEL: nAku Iwa vaccu.

'me-DAT swimming come'
TAM: eVnYakku nIccal weVriyum.
'me-DAT swimming know-fut-3p.sg.n'
ENG: I know to swim.

Handling Strategy :
Transfer Rule :
{[N1ku][N20][vaccu]}=>{[N1ku][N20][teri<3p.sg.n>]}

IV.b. TEL: AmeVku kadupu vacciMxi.
'She-DAT bellyt come-PST-3p.sg.n'
TAM: avalYukku karppam erYpattawu.
She pregnancy form-PRS-3p.sg.f'
ENG: She became pregnant

In the above example, the idiomatic sense of 'kadupu' is 'pregnant' which means 'the stomach'. But Tamil uses the term 'karppam' to express the same.

4. Conclusion: The Telugu- Tamil Machine Translation system is built by using the concept based dictionaries discussed above. The concept based dictionaries ensure the resolution of much of the disambiguation presented by the words in the lexical substitution in translation. The system is tested continuously by the native speaker of Tamil in order to validate its performance in the translation. The five scale Evaluation method of IL-IL MT is adopted for this purpose. The current comprehension of the outputs fall between **85-90**%.

[1] Transliteration Scheme using wx-notation:
Tamil Orthography :
a A i I u U eV e E oV o O H
k f c F t N w n p m y r l v lYY lY rY nY j s h R

Telugu Orthography :
a A i I u U q Q eV e E oV o O M H
k K g G f c C j J F t T d D N w W x X n p P b B m y r rY l lY lYY v S R s h

References:

[1] Arden A.H. 1891. *A Progressive Grammar of the Tamil Language*. Chennai: The CLS.
[2] Bhuvaneswari . G. 2009. *Telugu-Tamil Machine Transaltion.* Unpublished Ph.D. Thesis, University of Hyderabad.
[3] Dorr, Bonnie. 1990b. *Solving Thematic Divergence in Machine Translation*. In the Proceedings of the 28th Anual Conference of the ACL,127-134, University of Pittsburg, Pittsburg, PA.
[4] Dorr, Bonnie . 1993. *Machine Translation: A View from the Lexicon.* Cambridge, Mass: The MIT Press.
[5] Krishnamurti, Bh and Gwynn, J.P.L. 1985. *A grammar of modern Telugu*. New Delhi: OUP.
[6] Mohanty Rajat K., Bhattacharya P and et. al. *Synset Based Multilingual Dictionary : Insights, Applications and Challenges.* : www.cse.iitb.ac.in / ~pb/papers/ gwc08- **multilingual- dictionary**.pdf
[7] Sangal Rajeev, Uma Maheshwar Rao G, Nagamma Reddy K, 1999. *preceedings of the National Seminar of 'information Revolution and Indian Languages'* , Society for Computer Applications in Indian Languages : Hyderabad.
[8] Sinha, R.M.K., Thakur, A. 2005. *Translation Divergence in English-Hindi MT EAMT*, Budapest, Hungary.
[9] Uma Maheshwar Rao, G. 2002. *A Computational Grammar of Telugu*. (Momeo). Hyderabad: University of Hyderabad.