Development of Telugu-Tamil Transfer-Based Machine Translation system: With Special reference to Divergence Index

K. Parameswari
Centre for Applied Linguistics and Translation Studies
University of Hyderabad
pksh@uohyd.ernet.in, parameshkrishnaa@gmail.com

Abstract

The existence of translation divergence precludes straightforward mapping in machine translation (MT) system. An increase in the number of divergences also increases the complexity, especially in linguistically motivated transfer-based MT systems. In other words, divergence is directly proportional to the complexity of MT. Here we propose a divergence index (DI) to quantify the number of parametric variations between languages, which helps in improving the success rate of MT. This paper deals with how to build divergence index for a given language pair by giving examples between Telugu and Tamil, the major Dravidian languages spoken in South India. It also proposes handling strategies to overcome these divergences. The presentation of the paper also includes a live demo of Telugu-Tamil MT.

1 Introduction

In MT, there are a number of methods that are being practiced all over the world, chiefly, they are direct, interlingual, transfer-based methods and a combination of these beside the statistical and corpus based methods. This paper discusses the development of transfer-based Telugu-Tamil MT system with a special reference to divergences. In the development of MT[1], linguistically-grounded classification of divergence types need to be formally defined and systematically resolved. Identifying such divergences is the most significant part that facilitates the design and implementation of MT systems. As divergences are encountered as the specific problem in MT, identifying these are also the most crucial to obtain qualitatively a better output.

Divergence between languages may vary from one language pair to another. An increase in the number of divergence also increases the complexity in building an MT. In other words, it can be stated that divergence is directly proportional to the complexity of MT. Measuring divergence between languages supports to ascertain effort justification to build an MT for the proposed languages. Here we propose a divergence index (DI) to quantify the number of parametric variations between languages. DI also classifies divergence exhaustively into different levels in order to understand its depth. It facilitates MT in proposing where to put efforts for the given language pair to attain a better result.

2 Telugu-Tamil MT

Telugu and Tamil are major Dravidian languages with rich literary tradition sharing indubitable linguistic similarities and dissimilarities. An MT between them may be viewed as a bridge to understand and share the richness of both the languages. The MT system demonstrated here is a completely automatic translation system without human interference for the first time involving Telugu and Tamil. It is one of the successfully implemented systems under Indian language to Indian language(IL-IL) MT[2].

[1] In this paper, MT refers to linguistically motivated transfer-based machine translation.

[2] IL-IL MT is a consortium funded by the Department Of Electronics and Information Technology (DeitY), Ministry Of Communications and Information Technology, Government of India under the project name *Sampark*. Telugu-Tamil MT is

Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015), pages 48–54,
Praha, Czech Republic, 3–4 September 2015.

The Telugu-Tamil MT system is an assembly of various linguistic modules run on specific engines whose output is sequentially maneuvered and modified by a series of modules till the output is generated. It employs three stage architecture:

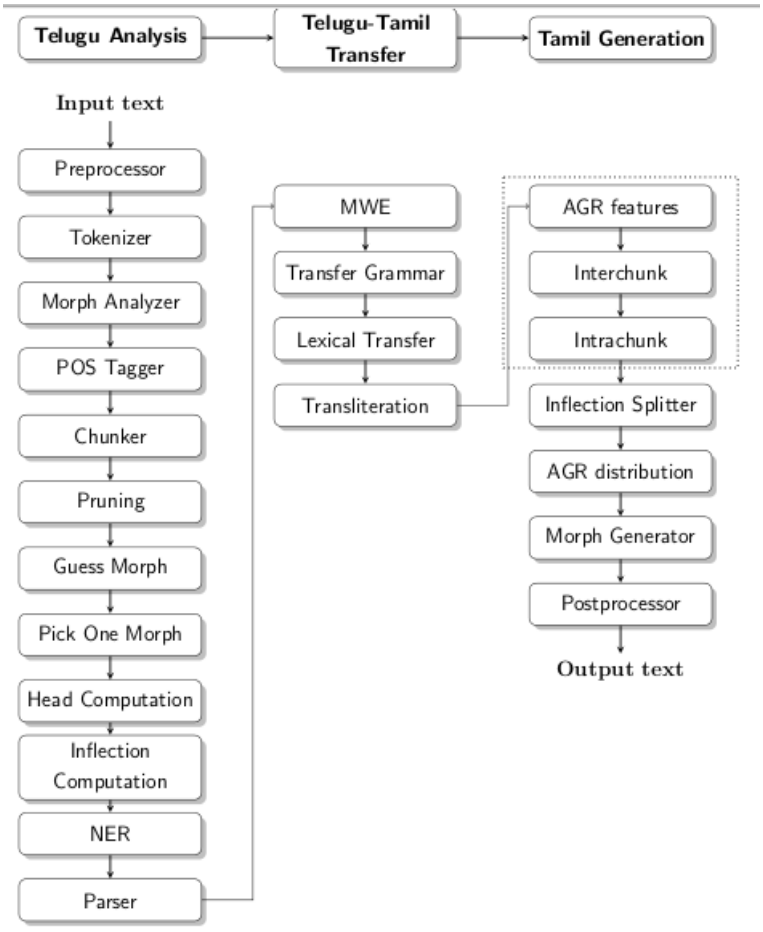
Stage 1: Source language analysis

Stage 2: Source language to target language transfer

Stage 3: Target language generation

The most crucial linguistic modules in Source Language analysis include a Morphological Analyzer (MA), Parts of Speech Tagger (POS), Chunker, Named Entity Recognizer (NER), Simple Parser (SP) and the Source Language to Target Language Transfer Module includes Multi Word Expression (MWE) component, Transfer Grammar (TG) Component, Lexical Transfer component consisting of a synset and bilingual lexicons and in Target Language Generation includes Agreement (AGR) modules and a Morphological Generator (MG). All the modules have been integrated on the platform called *Dashboard* based on black board architecture (Pawan et al, 2010) which configures data flow in a specified pipeline.

The architecture of Telugu-Tamil MT system is given below:



3 Translation Divergence

The term 'Translation Divergence' refers to distinctions or differences that occur between languages when they are translated. It is realized when the source language content is decoded differently in the target language and affects the 'well-formedness' of the target language. According to Dorr (1993), the translation divergence occurs when the underlying concept or 'gist' of a sentence is distributed over dif-

being developed as part of a larger project of IL-ILMT (with Prof. G. Uma Maheshwar Rao as the chief investigator) at Language Technology Laboratory, Centre for Applied Linguistics and Tanslation Studies(CALTS), University of Hyderabad. For more details see: http://caltslab.uohyd.ernet.in/. This system is also available at TDIL website for public access: http://tdil-c.in/components/com_mtsystem/CommonUI/homeMT.php.

ferent words or in different configurations for different languages. The notion of divergence in MT is comparable to the linguistically motivated notion of *parameteric variation* i.e. cross-linguistic distinctions.

Telugu and Tamil in spite of being cognate languages, exhibit considerable amount of divergences in various levels affecting the quality of output. In most of the cases, Dorr's divergences are noticed as rare phenomena and do not pose much problem as far as Telugu and Tamil are considered. However, these language pairs do pose problems at various other levels displaying different divergences. The current research attempts to classify these divergences into three major kinds, such as morphological, syntactic and lexical-semantic divergences.

4 Divergence Index

Divergence index (DI) represents a measure of the differences that occur between languages. The variations of linguistic features can be seen at any levels (L) in terms of surface, shallow and deep levels of languages. These levels are identified as L1, L2, L3 etc., according to its depth of variation. Identifying the divergence with its level between a pair of languages enables one to compute and quantify the effort that is required to build an MT. DI uses a table that attributes to features to identify and classify divergences exhaustively into different levels in order to understand its depth. It facilitates MT in proposing where to put effort for the given language pair to obtain a better result.

4.1 Divergence Index Table

Languages may share certain features or differ with each other. When they differ, it indicates that a certain feature is encoded differently or not available in one of the languages. This is a cause for divergence. Table 1 provides instances where divergences are possible with reference to a given feature in the said languages. Y indicates that the feature is available in a language and N indicates not. When both the languages share similar features (see Table 1 (1.) and (2.)), it means no divergence (indicated by 0). When they differ (see Table 1 (3.) and (4.)), there arises divergence (indicate by 1).

S.No	SL feature	TL feature	Divergence Index
1.	Y	Y	0
2.	N	N	0
3.	Y	N	1
4.	N	Y	1
5.	Y/N	Y/N	0
6.	Y/N	Y	1
7.	Y/N	N	1
8.	Y	Y/N	0
9.	N	Y/N	0

Table 1: **Divergence Index Table**

In certain cases, Y/N is given to indicate optional in the use of a feature. When both source language (SL) and target language (TL) show optional, it means no divergence (see Table 1 (5.)). When only SL shows optional, it is counted as divergence because TL element may not be directly mapped when the option differs (see Table 1 (6.) and (7.)). When the option occurs only in TL, it is counted as no divergence (see Table 1 (8.) and (9.)) because TL optionally behaves like SL, hence SL features can be directly mapped to TL.

4.2 Morphological Divergence Index

Morphological divergences, here, we refer to divergences that occur due to inflectional and productive derivational devices of words between Telugu and Tamil. Open word class categories such as nouns,

verbs and adjectives and closed word classes such as pronouns, number words and nouns of space and time (NST) are studied to find out morphological divergences. Functional elements on these categories need to be carefully matched from the source language to the target language to attain well-formed wordforms in the output. Uninflected word classes i.e. indeclinables and non-productive derivational wordforms are excluded here because they are listed in the lexicon and straightforward mapping between them solves the problem in MT.

For example , nouns in Telugu and Tamil are major word classes inflecting for number and case. The major inflectional differences occur due to two reasons i.e. (1) the choice of items in terms of inflections viz., the oblique stem formation, case and postposition and (2) the order of their presentation. For instance, the Table 2 explicates the differences.

No.	PSP	Telugu	Tamil	Gloss
1.	Comparitive	*iMṭi- kaMṭē* house.OBL- than	*vīṭṭ- ai- viṭa/* *vīṭṭ- ai.k- kāṭṭilum* house- ACC- than	'compared to the house'
2.	Semblative	*iMṭi- lāMṭi/* *iMṭi- vaMṭi/* house.OBL- like *iMṭi- ni- pōlina* house- ACC- like	*vīṭṭ- ai.p- pōṉṟa* house- ACC- like	'like the house' (adnominal usage)
3.	Locative: Circumferential	*iMṭi- cuṭṭū* *iMṭi- cuṭṭūtā* house.OBL- around	*vīṭṭ- ai.c- cuṟṟiyum/* *vīṭṭ- ai.c- cuṟṟilum* house- ACC- around	'around the house'
4.	Locative:Interior:Direction	*iMṭi- lōpali- ki/* *iMṭi- lō- ki* home.OBL- inside- DAT	*vīṭṭ- ukk- uḷ(ḷē)/* home.OBL- DAT- inside *vīṭṭ- iṉ- uḷ(ḷē)* home.OBL- GEN- inside	'to inside the house'

Table 2: **Postpositions**

As seen in the table 2 (No. 1-3), certain postpositions require their complement nouns differently case marked between Telugu and Tamil. Also as shown in Table 2 (No. 4) the order of suffixes in Telugu and Tamil may differ. The difference is explicated as below:

Te. Noun- ±Number suffix- ±Stem-formative- ±Postposition- ±Case Suffix

Ta. Noun- ±Number suffix- ±Stem-formative- ±Case Suffix- ±Postposition

The divergence index for Table 2 is built as below:

No.	PSP	Telugu	Tamil	DI/Level
1.	Comparative	Y	Y	0/L1
	Accusative case marker	N	Y	1/L2
2.	Semblative	Y	Y	0/L1
	Accusative case marker	Y/N	Y	1/L2
3.	Locative: Circumferential	Y	Y	0/L1
	Accusative case marker	N	Y	1/L2
4.	Locative:Interior:Direction	Y	Y	0/L1
	Dative case marker with PSP	Y	N	1/L2

Table 3: **Divergence Index for Table 2**

In predicative positions, nouns in Telugu agree with their subjects in the first person singular and plural, and in the second person singular and exhibit explicit overt markings unlike Tamil. Consider the following in Table 4.

S.No.	GNP	Telugu	Tamil	Gloss	DI/Level
1.	1.SG.	*maniṣi-* ***ni*** human.SG.OBL-**1.SG.**	*maṉitaṉ-* ***ø*** human.SG	'(I am) a human'	1/L1
2.	1.PL.	*manuṣula-* ***mu*** human.PL.OBL-**1.PL.**	*maṉitar- kaḷ-* ***ø*** human- PL	'(we are) humans'	1/L1
3.	2.SG.	*maniṣi-* ***vi***	*maṉitaṉ-* ***ø***	'(You are) a human'	1/L1

	human.SG.OBL-**2.SG.**	human.SG

Table 4: **Nominal predicates in Telugu and Tamil**

These kind of divergences need to be noticed and handled strategically in the target language Tamil since it does not express these details on nominal predicates. Morphological divergences are mainly handled by the morphological generator (MG), the target language (TL) generation module. MG is equipped with inbuilt morphological features of TL which generates acceptable TL. Other modules such as parser, transfer grammar (TG), lexical transfer (LT) and agreement (AGR) modules do involve in handling morphological divergence.

4.3 Syntactic Divergence Index

Syntactic divergence here we refer to syntactic structural differences that occur between pairs of languages. It is obvious to find out similar constructions in Telugu and Tamil in majority of cases but still there are lots of variations arise due to case mismatches, agreement, anaphora, negation, subordination and clitics. Various syntactic processing and a robust transfer grammar are obviously required to overcome syntactic divergence.

For example, each case marker has a number of functions and it is obvious that they lead to case mismatches in MT. The difference in form and function of a case in the source language precludes the straightforward mapping of it in the target language. For instance, Telugu and Tamil agree in using the dative case marker in various functions viz., beneficiary of an action, goal of motion, experiencer subject (Cf. Krishnamurti, 2003:434; Verma and Mohanan, 1990:27) among other functions. However, to express a possessive relationship between two inanimate nouns, one of the nouns of inanimate category carries the dative marker to express the locative function in Telugu. On the contrary, the locative case marker is in use in Tamil. Example:

Te. *gōḍa- ku kiṭikī uM- di.*
wall- DAT window.NOM be.PRS- 3.SG.N.

Ta. *cuvarr- il jannal iru- kkir- atu.*
wall- LOC window.NOM be- PRS- 3.SG.N.
'The wall has a window.'

Syntactic divergences are mainly handled by TG. TG is equipped with performing certain tasks such as insertion, deletion, modification and re-ordering of words and chunks. It also has the ability to handle files where it is possible to operate a single rule over a list of items.

4.4 Lexical-Semantic Divergence Index

Lexical-semantic translation divergences are characterized by properties that are entirely lexically determined between languages. A concept expressed by a lexeme may not have the similar meaning in all contexts. The major lexical-semantic divergences that occur between Telugu and Tamil are due to the nature of its semantic compositions and their formal collocation in their expression.

For example, a lexeme, used to express a concept in a language may not have the same meaning in all contexts. When it has multiple meanings, word sense disambiguation is required to overcome lexical ambiguity and to select an appropriate sense with its form in the target language.

For instance, the lexeme *kuṭṭu* in Telugu is ambiguous and expresses three different senses as given below:
Sense 1: *kuṭṭu* 'to bite' as in the context of *cīma* 'an ant' and etc. The equivalent word in Tamil is *kaṭi* 'to bite'.
Sense 2: *kuṭṭu* 'to stitch' as in the context of *baṭṭalu* 'clothings'. The equivalent word in Tamil is *tai* 'to stitch'.

Sense 3: *kuṭṭu* 'to pierce' as in the context of *cevulu* 'ears' or body parts and etc. The equivalent word in Tamil is *kuttu* 'to pierce'.

Lexical-semantic divergences are handled by MWE component and LT. MWE component contains a lexical database consisting words of co-occurrence. When a group of words are identified as MWE, this module transfers them into the acceptable target language expression. Lexical ambiguities are handled by TG. An exhaustive set of transfer grammar rules operating on identification of the ambiguous words and disambiguating them by looking at the subject or the object nouns as suggested above are built. For instance, the following TG rules are samples to handle the different senses of Telugu word *kuṭṭu* in Tamil.

```
V1:R1::"$x=animate.txt"
R1:  NP<root="$x",lcat="n"> VGF<root="kuṭṭu",lcat="v"> =>
NP<root="$x", lcat="n"> VGF<root="kaṭi",lcat="v">
V2:R2::"$y=inanimate.txt"
R2:  NP<root="$y",lcat="n"> VGF<root="kuṭṭu",lcat="v"> =>
NP<root="$y", lcat="n"> VGF<root="tai",lcat="v">
V3:R3::"$z=bodyparts.txt"
R3:  NP<root="$z",lcat="n"> VGF<root="kuṭṭu",lcat="v"> =>
NP<root="$z", lcat="n"> VGF<root="kuttu",lcat="v">
```

5 Conclusion

Though Tamil and Telugu belong to the same language family (Dravidian language family), some major and minor differences are found in their linguistic behavior which preclude any straightforward mapping. To avoid this, it is essential to formalize the divergent patterns and develop a certain number of rules as the case demands to have a successful system with broad coverage. Building divergence Index is proved to be a useful activity to identify and handle divergences effectively in transfer-based MT.

References

Annamalai, E. 2000. 'Lexical Anaphors and Pronouns in Tamil'. In Lust et al (ed.), *Lexical Anaphors and Pronouns in Selected South Asian Languages: A Principled Typology*, 169–216.

Arokianathan, S. 1981. *Tamil Clitics*. Trivandrum: Dravidian Linguistics Association.

Bharati, Akshar, Rajeev Sangal & Dipti M Sharma. 2007. 'SSF: Shakti Standard Format Guide' 1–25.

Dash, Niladri Sekhar. 2013. 'Linguistic Divergences in English to Bengali Translation'. *International Journal of English Linguistics* 3(1).

Dave, Shachi, Jignashu Parikh & Pushpak Bhattacharyya. 2001. 'Interlingua-based English–Hindi Machine Translation and Language Divergence'. *Machine Translation* 16(4). 251–304.

Dorr, Bonnie Jean. 1993. *Machine Translation: a View from the Lexicon*. Massachusetts: MIT press.

Dorr, Bonnie Jean. 1994. 'Classification of Machine Translation Divergence and a Proposed Solution'. *Computational Linguistics* 20(4). 597–633.

Emeneau, Murray B. 1956. 'India as a Linguistic Area'. *Language* 3–16.

Goyal, Pawan & R Mahesh K Sinha. 2009. 'Translation divergence in English-Sanskrit-Hindi Language Pairs'. In *Sanskrit Computational Linguistics*, vol. 5406, 134–143. Springer.

Gupta, Deepa & Niladri Chatterjee. 2003. 'Identification of Divergence for English to Hindi EBMT. In *Proceeding of MT Summit-IX*, 141–148.

Hockett, Charles F. 1954. 'Two Models of Grammatical Description'. *Word 10.* 210–234.

Krishnamurti, Bh. & J. P. L. Gwynn. 1985. *A Grammar of Modern Telugu*. Delhi: Oxford University Press.

Lehmann, Thomas. 1993. *A Grammar of Modern Tamil*. Pondicherry: Pondicherry Institute of Linguistics and Culture.

Masica, Colin P. 1976. *'Defining a Linguistic Area: South Asia'*. Chicago: University of Chicago Press, 1976.

Mishra, Vimal & R. B. Mishra. 2008. 'Study of Example Based English to Sanskrit Machine Translation'. *Polibits* (37). 43–54.

Mitkov, Ruslan. 1999. *Anaphora Resolution: the State of the Art. http://clg.wlv.ac.uk/papers/mitkov-99a.pdf.*

Pawan, Kumar, A. K. Rathaur, Ahmad Rashid, K Sinha Mukul & Sangal Rajeev. 2010. 'Dashboard: An Integration & Testing Platform Based on Black Board Architecture for NLP Applications'. *Proceedings of 6th International Conference on Natural language Processing and Knowledge Engineering (NLP-KE)*, Beijing, China, August.

Shukla, Preeti, Devanand Shukl & Amba Kulkarni. 2010. Vibhakti Divergence between Sanskrit and Hindi. In *Proceedings of the International Sanskrit Computational Linguistics Symposium*, 198–208. Springer.

Subbarao, K. V. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge: Cambridge University Press.

Subbarao, K. V. & B. Lalitha Murthy. 2000. 'Lexical Anaphors and Pronouns in Telugu'. In Lust et al (ed.), *Lexical Anaphors and Pronouns in Selected South Asian Languages: A Principled Typology*, 217–276.

Verma, Manindra K. & Mohanan K. P. (eds.). 1990. '*Experiencer Subjects in South Asian Languages'*. Stanford: Center for the Study of Language (CSLI).

Whitman, Neal. 2002. 'A Categorial Treatment of Adverbial Nouns'. *Journal of Linguistics* 38. 561–597.

Weaver, Warren. 1955. 'Translation'. *Machine Translation of Languages* 14. 15–23.