A Transfer-rule Based Verb Phrase Translation from English to Tamil

Parameswari K.[1], Nagaraju V.[2], and Angeline Linda K.[1]

[1] University of Hyderabad
[2] eBhasha Setu Language Services
{parameshkrishnaa, vpp.nagaraju1234,angelineal1996}@gmail.com

Abstract. Building a machine translation (MT) between non-cognate languages always poses number issues as there are lots of translation divergences involved. In transfer-based MT, a systematic way of formulating transfer rules are required to handle linguistic differences between languages. This paper explains three-stages in which the transfer-based machine translation (MT) are built for translating verb phrases from English to Tamil.

Keywords: machine translation · divergence · verb phrase · transfer rules

1 Introduction

Machine translation is one of the challenging tasks in NLP as it involves a deep understanding of the source text and generating the acceptable target language automatically. Further, translation between non-cognate languages requires more efforts as there are more divergences i.e. cross-linguistic differences [1] which affects the well-formedness of the target-language generation. In this paper, an effort towards building transfer rules for automatic verb phrase translation (VPT) from English to Tamil is attempted. This paper explains three-stages in which the transfer-based machine translation (MT) are built for translating verb phrases from English to Tamil.

2 Linguistic Typology of Verb Phrases: English and Tamil

The configuration of verb phrases in English and Tamil show lots of differences which precludes the straightforward mapping of lexical and grammatical elements between these languages. Tamil is known for agglutinating morphology and encodes various grammatical information as suffixes on verbs. Whereas English is known as morphologically poor language, hence the grammatical information is realized as different words in strict word-order.

This major linguistic typological differences of verb phrases in English and Tamil are listed below:

- The finite structure of the verb phrase in English and Tamil is:
  English: Model+Auxiliary/-ies+ Main Verb
  Tamil: Main Verb + Auxiliary/-ies+ Model+ Person-Number-Gender (PNG) Agreement
- Negation is expressed through inflection on verbs in Tamil whereas through auxiliary in English.
- In Tamil, the copula verb *āku* 'to be' optionally occurs in the sentences of nominal predicates, whereas the copula is obligatory in English.
- Compound verbs are constructed where sequences of a 'polar' verb followed by one or more of 'vector' verbs in Tamil. The polar verb is realized in Verbal participle or infinitive form in Tamil, unlike English.
- In Causative construction, the periphrastic causative auxiliary verb *-vai* occurs in Tamil. Whereas the causative verb precedes the main verb in English.
- Reflexive and reciprocity are expressed through the auxiliary verb *-koḷ* and optionally through reflexive pronouns in Tamil. Whereas, English uses pronouns to express the same.
- Conjunct verbs (noun plus light verb) are more productive in Tamil when compared to English.
- Non-finite verbs which head the subordinate clause inflects for verbal participle, infinitive, conditional and concessive forms in Tamil, unlike English.
- In relative clause construction, Unlike English, Verbs in its adjectival participle form occurs before the noun phrase in Tamil.
- Clitics such as interrogative, dubitative, emphatic and inclusive markers are added with verbs in the end position in Tamil, unlike English where these markers are expressed by different elements.

3 English-Tamil MT: A Review

Number of activities in building MT between English-Tamil are attempted by various groups and researchers in India. It includes Anuvadaksh (English to Indian Language Machine Translation System), Soman and Menon et.al [9], Poornima et.al [5], Saravanan [8], Pandian and Kathirvel [3], Ramaswamy et.al [7], Kumar et.al [2], Rajeswari et.al [6] to name a few. In this paper, an attempt is made in building transfer-based approach to MT between English-Tamil.

4 Algorithm for Verb Phrase Translation

This section reports the algorithm used in verb phrase translation from English to Tamil. The algorithm used in building VPT consists of three stages:

4.1 Identification of Verb Phrase (VP) and its subject (nsubj)

This stage identifies the verb phrase from the shared map file and also identifies the subject(s) (nsubj) of the sentence using the dependency-based parser. The nsubj is identified for their PNG features and the same is percolated to the verb to which it is identified as nsubj. The Tense (T) information is retrieved from POS of the head verb. The algorithm is given below:

```
1: Get Input Sentence
2: Identify Verb Phrase (VP) from shared Mapfile
3: Call Parser ( nltk.parse.stanford)
4: Find ‘nsubj’ form Parse output
   match nsubj with NP list (consisting GNP features)
     If found
         get PNG feature
     else
         add default PNG feature (3,sg,n)
5: Find head verb of VP and its nsubj from Parse output
     If found
         percolate nsubj PNG features to head verb
     else
         add default PNG feature (3,sg,n) to head verb
6: Identify T based on POS output of Parse
```

Algorithm 1: Identification of VP and nsubj

4.2 Stage 2: Transfer Rules

This stage involves transferring the structure and lexical items of English to Tamil. Using nltk lemmatizer, the verb root ‘VR’ is identified. The structure of verb phrases is transferred to Tamil based on Rules (28 rules as an initial attempt) that are compiled in mapping English-Tamil. The lexical substitution from English to Tamil is executed in this stage. The algorithm is given below:

```
1: Get head verb
2: Identify verb root (VR) (lmtzr.lemmatize)
3: check head verb co-occurrence
(e.g. Auxiliary verbs, Model verbs, Negation, Wh-question etc.,)
     if found
         Apply Rules to reorder them
4: Identify Main and Subordinate clauses from Parse cues
     if found a subordinate clause
         Apply Rules to provide appropriate TAM
5: match VR in English to Tamil verb dictionary
     if found
        Substitute VR with equivalent Tamil
    else
         Transliterate English VR into Tamil
```

Algorithm 2: Transfer Rule Application

4.3 Stage 3: Generation

Generating well-formed wordforms based on identified PNG and TAM features in Tamil is attempted in this stage. The morphological generator for Tamil [4] is used for generating the Tamil verbs. The algorithm is given below:

```
1: Get Tamil VR, PNG and TAM features
2: Call Tamil Morphological Generator (TMG)
3: Input ‘VR,lcat (v), G, N, P,TAM’ to TMG
4: Get the output
```

Algorithm 3: Generation

5 Experiments and Results

The current MT system is evaluated by the coordinators of shared task on verb phrase translation in English and Indian Languages (VPTIL). The total number of training VPs in sentences received are 2275 and the total of testing VPs in sentences are 1869. The scoring criteria (see Table 1) and results (see Table 2) obtained are given below:

Table 1. The Scoring Criteria

Completely Correct	Score 4
TAM and PNG Correct	Score 3
Correct root and TAM partially correct	Score 2
Correct root and wrong TAM	Score 1
Completely Incorrect	Score 0

Table 2. Results of English-Tamil MT

Precision	20.77%
Recall	28.95%

6 Conclusion

The system performance can be improved by the improvement in transfer-rules, source language analysis modules, and target language generation modules. A robust lexical substitution is also required for the effective mapping of source language verb root to the target language of the system. The precision of the system reveals that the transfer rule-based approach to MT between English-Tamil performs well and can be improved further with the addition of new rules.

Abbreviations: VP- Verb Phrase; nsubj- Subject of verb phrase; PNG Person-Number-Gender markers; POS- parts-of-speech tagging; TAM- tense, aspect and model marker(s); lcat- lexical category; TMG- Tamil Morphological Generator; VR- verb root

Acknowledgment: The team acknowledges the coordinators of shared-task of VPTIL for their inputs.

References

1. Dorr, Bonnie Jean. *Machine Translation: a View from the Lexicon.* Massachusetts: MIT press. (1993)
2. Kumar A.M., V. Dhanalakshmi, K. P. Soman and S. Rajendran, Factored statistical machine translation system for English to Tamil language, *Pertanika J. Soc. Sci. Hum.* 22 1045–1061 (2014).
3. Pandian L.S. K. Kadhirvelu, Machine translation from English to Tamil using hybrid technique, *Int. J. Comput. Appl.* 46 (2012).
4. Parameswari, K. An Implementation of Apertium Morphological Analyzer and Generator for Tamil. *Language in India* 11. 71–75 (2011).
5. Poornima C., V. Dhanalakshmi, M. Anand Kumar K. P. Soman, Rule based sentence simplification for English to Tamil machine translation system, *Int. J. Comput. Appl.* 25 (2011).
6. Rajeswari S., P. Sethuraman K. Krishnakumar, English to Tamil machine translation system using universal networking language, *Sādhana* 41 607–620 (2016).
7. Ramasamy L., O. Bojar and Z. abokrtsk, Morphological processing for English-Tamil statistical machine translation, in: *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pp. 113–122, (2012).
8. Saravanan S., *English to Tamil machine translation: rule based approach,* LAP LAMBERT Academic Publishing, (2012).
9. Soman, K.P. A. G. Menon, English to Tamil machine translation system, in: 9th Tamil Internet Conference (INFITT), Chemmozhi Maanaadu, Coimbatore, India, (2010).