A vis-à-vis evaluation of MT paradigms for linguistically distant languages

Ruchit Agrawal
LTRC
IIIT Hyderabad

Jahfar Ali
LTRC
IIIT Hyderabad

Dipti Misra Sharma
LTRC
IIIT Hyderabad

Abstract

Neural Machine Translation is emerging as the de facto standard for Machine Translation across the globe. Statistical Machine Translation has been the state-of-the-art for translation among Indian languages. This paper probes into the effectiveness of NMT for Indian languages and compares the strengths and weaknesses of NMT with SMT through a vis-a-vis qualitative estimation on different linguistic parameters. We compare the outputs of both models for the languages English, Malayalam and Hindi; and test them on various linguistic parameters. We conclude that NMT works better in most of the settings, however there is still immense scope for the betterment of accuracy for translation of Indian Languages. We describe the challenges faces especially when dealing with languages from different language families.

1 Introduction and Related Work

There is an immense scope in the development of translation systems which cater to the specific characteristics of languages under consideration. Indian languages are not an exception to this, however, they add certain specifications which need to be considered carefully for effective translation. Firstly, they span across multiple language families like the Indo-Aryan and Dravidian languages. Secondly, there is a lack of large parallel corpora for most of these languages, which are required to build robust systems by the SMT and NMT paradigms.

This paper probes in to the competence of different MT paradigms with respect to language pairs which belong to different language families. Dravidian languages raise many intriguing issues in modern linguistics. One of them is the differentiation of the finiteness and non finiteness of clauses with its tense inflection in verbs (Amritavalli, 2014), (McFadden and Sundaresan, 2014), (Tonhauser, 2015). Scrambling effect on canonical word order (Jayaseelan, 2001) is another such phenomenon. It is to be observed when dealing with complex syntactical structures containing cleft constructions in Malayalam (Jayaseelan and Amritavalli, 2005).

Relative clause structures, nominal clause structures and their coordination constructions in Dravidian languages are other interesting phenomena (Amritavalli, 2017), (Jayaseelan, 2014). The analysis made in the paper describes the handling of all these linguistic phenomena in the context of machine translation.

Neural Machine Translation is emerging as a de facto standard for Machine Translation across the globe. However, a manual inspection of the output translations reveals significant scope for improvement in translation quality. We perform a comparative analysis of Neural and Phrase-based Statistical MT techniques and highlight the strengths and weaknesses of each paradigm with respect to handling of different linguistic phenomena. The enquiry throws light on some of the challenging cases encountered when translating between morphologically rich and free word order languages and the other end of morphologically less complicated and word order specific languages. A set of basic observations are made after extensive testing of SMT and NMT outputs on these language pairs. We observe that NMT performs better than SMT for most of the linguistic phenomena considered. However; one of the major hurdles to

Proc. of ICON-2017, Kolkata, India. December 2017 ©*2017 NLPAI*, pages 33–42

deliver the correct output between morphologically rich languages to to morphologically weak languages is the inadequacy of NMT to generate word forms with correct affixes.

The analysis can generate fruitful insights in the modification of NMT / SMT based techniques to generate efficient results. The insights can be taken into consideration during the building of parallel corpora in the future or using linguistic features as additional informaiton for training NMT models. The analysis also enables the usage of a particular paradigm depending upon the language pair and domain in consideration.

2 Motivation

2.1 Characteristics of Indian languages

The majority of Indian languages are morphologically rich and depict unique characteristics, which are significantly different from languages such as English. Some of these characteristics are the relatively free word-order with a tendency towards the Subject-Object-Verb (SOV) construction, a high degree of inflection, usage of reduplication, conjunct verbs, relative participal forms and correlative clause constructions. These unique characteristics coupled with the caveats of evaluation metrics described in Section 2.3 pose interesting challenges to the field of Indian Language MT - both in terms of development of efficient systems as well as their evaluation.

For example, in Hindi, a sentence s containing the words w_1,w_2,..,w_n can be formulated with multiple variants of word ordering. This behavior is depicted in Table 1, which shows two Hindi translations of the following English sentence :
'Shyam has given the book to Manish.' Although they use different word-order, both of them are semantically equivalent and correct translations of the source sentence.
Similarly, for the sentence 'The sun has set', there can be multiple valid translations, as shown in Table 2. It can be noted that 'सूर्य ' and 'सूरज ' are synonyms of 'Sun' in Hindi.

In addition to these, there are many subtle differences in the ways different Indian languages encode information. For example, Hindi has two genders for nouns whereas Gujarati has three. There are also many ambiguities introduced in language (both at lexical as well as sentence levels) due to the socio-cultural reasons and partial encoding of information in discourse scenario. In addition to this, the majority of Indian languages encode a significant amount of linguistic information in their rich morphological structures, and often lexemes can have multiple senses. All these factors like linguistic conventions, socio-cultural knowledge, context and highly inflectional morphology combined together make Indian languages a challenging terrain for Machine Translation.

2.2 Variation in linguistic constructions in IA and DR languages

Even though Indian languages are all typologically SOV, there are distinct syntactic peculiarities in Dravidian languages (DR) that makes MT challenging between Indo-Aryan (IA) and Dravidian languages. Two such phenomena are shown by the examples below:

1.
   - Hindi Sentence : राम ने बोला कि वह घर जा रहा था
   - Transliteration : rām nē bōl-ā ki vah ghar jā rahā thā
   - Gloss : Ram ERG tell-PST S.CONJ 3.SG.D.PRON home go AUX1-CONT AUX2-PST
   - Meaning : Ram said that he is going home

2.
   - Malayalam Sentence : അവൻ വീട്ടിലോക്ക് പോകുകയാണ് എന്ന് രാമൻ പറഞ്ഞു
   - Transliteration : avan vīṭṭilēākk pēākukayāṇ enn rāman paṟaññu
   - Gloss : He-NOM home-LOC-towards go-INF COP QT Raman-NOM say-PST.

3.
   - Telugu Sentence : రాముడు తాను ఇంటికి వెళ్తున్నట్టుగా చెప్పాడు

Table 1: Different Hindi translations corresponding to the English sentence - “Shyam has given the book to Manish.” (Due to word order)

	Hindi	Transliteration
Sent : 1	मनीष को श्याम ने किताब दे दी ।	maneesh ko shyaam ne kitaab de dee
Sent : 2	श्याम ने मनीष को किताब दे दी ।	shyaam ne maneesh ko kitaab de dee

Table 2: Two different translations corresponding to the English sentence - “The sun has set.” (Due to many-to-many mapping between vocabulary)

	Hindi	Transliteration
Sent : 1	सूर्य डूब चुका है ।	soorya doob chuka hai
Sent : 2	सूरज डूब चुका है ।	sooraj doob chuka hai

- Transliteration : rāmuḍu tānu iṇṭi-ki veḷ-tunn-aṭṭugā cepp-ā-ḍu
- Gloss : Ram 3P.REFL.PRON home-DAT go-PRES-MANNER.ADV tell-PST-3.M.SG

4.
- Tamil Sentence : ராமன் தான் வீட்டுக்கு செல்வதாக கூறினான்
- Transliteration : rāman̲ tān̲ vīṭṭu-kku cel-vat-āka kūṟ-in̲-ān̲
- Gloss : Ram 3P.REFL.PRON home-DAT go-NPST.R.PART-MANNER.ADV tell-PST-3.M.SG

Above example shows that in Hindi the main clause is followed by subordinate clause and both the clauses are connected by a subordinating conjunction ‘ki’. For Malayalam, The embedded clausal structure with quotative particle ‘ennu’ is the only kind of sentence possible to have two finite verbs (Asher and Kumari, 1997). In Telugu and Tamil (Dr), the subordinate clause is embedded within the main clause and connection between them is established morphologically through adverbial inflections or sometimes a quotative marker is used to connect the two clauses. These phenomena explain the relatively lower performance on Dravidian languages as compared to Indo-Aryan languages.

2.3 Challenges in automatic evaluation

A key aspect in developing efficient MT systems is addressing the issue of effective metrics for automatic evalution of translations, since manual evaluation is expensive and time-consuming. There has been significant interest in this area, both in terms of development as well as evaluation of MT metrics. The Workshop on Statistical Machine Transla- tion (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison- Burch et al., 2009) and the NIST Metrics for Machine Translation 2008 Evaluation 1 have both collected human judgement data to evaluate a wide spectrum of metrics. However, the problem of reordering has not been addressed much so far. The primary evalutaion metrics which exist currently for scoring translations are BLEU, METEOR, RIBES and NIST.

BLEU (Papineni et al., 2002) measures the number of overlapping n-grams in a given translation when compared to a reference translation, giving higher scores to sequential words. METEOR (Lavie and Denkowski, 2009) scores translations using alignments based on exact, stem, synonym, and paraphrase matches between words and phrases. RIBES (Isozaki et al., 2010) is based on rank correlation coefficients modified with precision. NIST (Doddington, 2002) is a variation of BLEU; where instead of treating all n-grams equally, weightage is given on how informative a particular n-gram is. We report the BLEU score as a measure to test accuracy for the 110 NMT systems to maintain brevity. However, for the language-pair English -> Hindi; we report all of the above scores. We also describe the challenges in evaluating MT accuracy keeping this language pair in consideration, however it should be noted that the same or similar challenges are faced when dealing with other language pairs as well. We use the

MT-Eval Toolkit[1] to calculate all these metrics.

It can be noted that most of the above-mentioned metrics employ some concept of word-order as well as word similarity using n-grams to score translations, which makes evaluating Hindi translations a tedious task. In addition to this, there exists a many-to-many mapping of vocabulary between English and Hindi which makes all of these scoring mechanisms less effective. For example, both translations shown in Table 2 are valid. However; since the current MT metrics rely heavily on lexical choice, there is no mechanism which takes into account the phenomena described above, which is which is quite common in Indic languages like Hindi. Hence, in addition to the metric scores, we also show sample examples with their descriptions in the following section, in order to demonstrate translation quality in a more comprehensive manner.

3 Parameters for evaluation

Since the evaluation metrics do not capture how well different linguistic phenomena are handled by our model, we perform a manual investigation and error analysis with the help of linguists. In order to have a clear insight of NMT performance as compared to SMT on various aspects, we do a side-by-side comparison of the output sentences generated by the SMT and the NMT models respectively. The linguists were asked to identify the strengths and weaknesses of NMT and SMT by ranking 200 output sentences produced by the respective models in terms of the following parameters:

- Word order
- Morphology :
  - How appropriate is the surface form selection
  - Usage of correct syntactic structures
  - Morphological agreement between words
- Phrase handling :
  - Non-translated phrases / phrases missing in the output
  - Additional phrases - Phrases occuring in the output but not in the input source sentence
- Lexical Choice - Quality and appropriateness of content words and terminology errors

[1] http://bit.ly/2p5C2FB

		Hindi	Malayalam	English
Hindi	SMT	-	10.4	27.87
	NMT	-	8.86	27.76
Malayalam	SMT	13.9	-	8.2
	NMT	12.56	-	7.88
English	SMT	26.84	5.15	-
	NMT	27.24	3.76	-

Table 3: Results of SMT and NMT on the ILCI test set

We show the results in Figure 1.

It can be observed from Figure 1 that SMT produces about twice as more errors in word order and almost thrice as more errors in syntactic and morphological structures and agreement than NMT. Thus the NMT model is able to perform significantly better than SMT for these phenomena. This results in much more fluent translations produced by the NMT model - making it a better choice in most scenarios. At the same time, the errors made in terms of lexical choice are much more in NMT than SMT. NMT also produces slightly greater number of errors in terms of missing or additional phrases. On deeper investigation, it is made clear that a majority of the lexical choice errors are due to the noise present in the training data. This leads to the insight that NMT is more prone to greater sensitivity to training noise than SMT.

To summarize, NMT performs better than SMT in most linguistic aspects, particularly in the presence of a high quality training corpus.

4 Analysis and insights

The analysis is based on the translation of prevalent sentence construction usages in the source languages. An extensive testing is done with these sentence constructions and some of the output has been reported with releveant translation and gloss in the

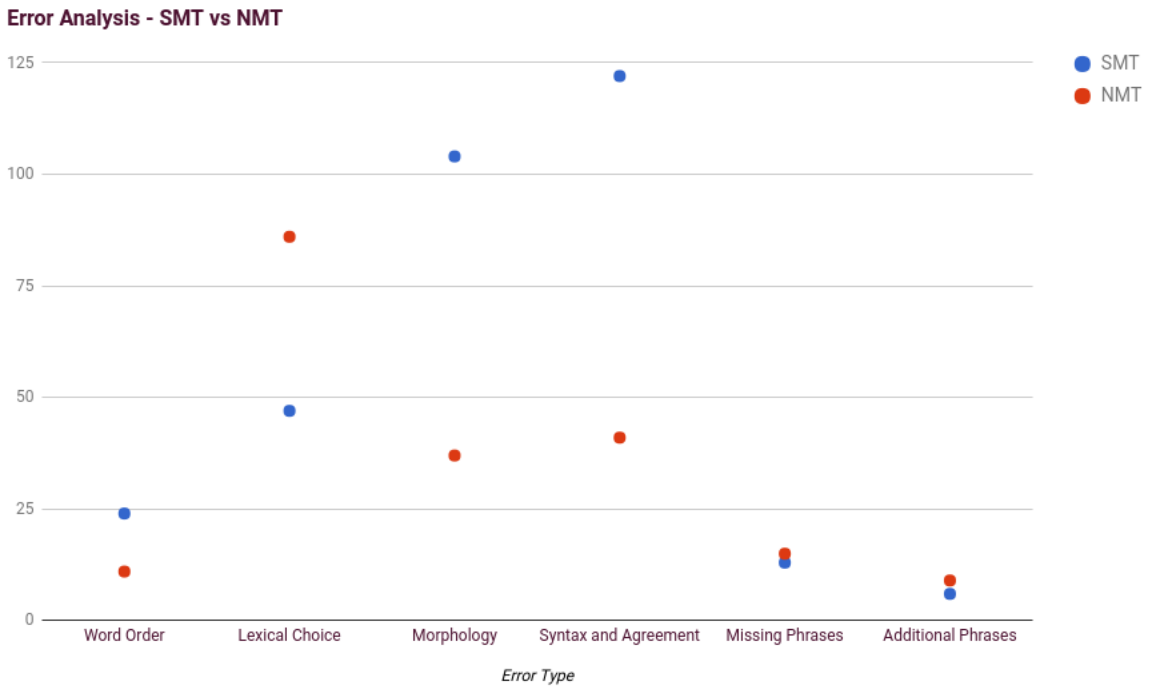

Figure 1: *Manual Error Analysis of performance of NMT with SMT*

coming sections. In order to understand the efficacy of capturing the syntactical structure of source language for the translation, we crafted simple sentences with different verbal inflections, such as transitive, intransitive, causative and different modalities in the source language. These sentences are tested and verified if the translated sentences are able to convey the same meaning from the source language. Similar attempt has been done with participle, cleft and coordination constructions in Malayalam to Hindi and English. For analysing Hindi to other languages, sentences with participle, complex predicate and coordination are tested to Malayalam and English. Tha analysis has done majorly on following sentences

- Simple sentences with different verbal inflections
- Participle, cleft and coordination constructions for Malayalam
- Participle, complex predicate, and coordination constructions for Hindi

4.1 Malayalam to English translation

SMT produces a lot of untranslated words as can be seen from the examples below, although the domain is kept the same for the manually created test set, however, the phrasing and structure is tweaked to cover all the grammatically possible constructions prevalent in Malayalam.

On the other hand, NMT shows an impressive performance in simple sentence translation from Malayalam to English. We observe that verbal inflections signalling modality is getting translated correctly in NMT(Example-3).Similarly NMT is also able to figure out variations in transitive and intransitive inflections in Malayalam to produce moderately equivalent English sentence(Example-2). At the same time NMT fails to translate imperative mood inflections correctly(Example-1). Example-1:
<SRC> വീട്ടില് കീടനാശിനി മരുന്നുകള് തളിക്കുക
<Gloss> home-LOC pesticide medicine spray-IMP.
<Translation> Pesticide sprays in home.
<SMT-ENG> വീട്ടില് കീടനാശിനി medicines തളിക്കുന്നു.

<Gloss> home-LOC pesticide medicine sprinkle-PRS.
<Translation> Pesticide medicine sprinkle in home.
<NMT-ENG> Get insecticides sprayed at home .
Example-2:
<SRC> വളരെ അധികം ദാഹം ഉണ്ടാകുന്നു
<Gloss> Very much thirsty make-PRS.
<Translation> It makes very much thirsty.
<SMT-ENG> feels very thirsty .
<NMT-ENG> One feels very thirsty .
Example-3: <SRC> ഉറങ്ങുന്ന സമയത്ത് കൊതുകു വല ഉപയോഗികണം
<Gloss> Sleep-RELAT time-DAT Mosquito net use-IMP.
<Translation> While sleeping mosquito net should be used.
<SMT-ENG> mosquito nets ഉപയോഗികണ while sleeping .
<Gloss> mosquito nets use-IMP. while sleeping .
<NMT-ENG> while sleeping mosquito net should be used.

4.1.1 Cleft constructions

Both paradigms fail to translate cleft constructions from Malayalam to English. Some of the complex syntactic constructions pertaining to the source or target language consistently fail to be learnt correctly , even though they are very common in the usage of the languages. The cleft construction could be accounted as an example as it is being used in both Malayalam and English. The SMT output is mostly erroneous and contains many untranslated words as can be seen from the following example.
Example-1:
<SRC> അനീമിക് സംബന്ധമായ രോഗങ്ങളെയാണ് വർദ്ധിപ്പിക്കുന്നത്
<Gloss> Anemic related-RELAT disease-COP increase-NOMIN.
<Translation> It is anemic relate diseases that are increased.
<SMT> anaemic വർദ്ധിപ്പിക്കുന്നു related diseases .
<Gloss> anaemic increase-PRS related diseases .
<NMT-ENG> It increases the diseases of anemic.

4.1.2 Participle constructions

The sentences with relative participle verb forms are translated incorrectly from Malayalam to English and from English to Malayalam as well. The relativised form of the verbs are predominently used in Malayalam for relative clause construction. It extends a subject sharing possibility between the realtive clause and the main clasue without the need of pronoun usage. It has also been observed that complex postpositional phrases and nomilised clauses are translated well in NMT in both directions. The example shows an erroneous traslation of a relative participle clause usage in the sentance.
Example-1:
<SRC>രോഗം പരത്തുന്ന കൊതുക് ഏഡിസ് എഡിപ്പടായ് ആണ്
<Gloss> Disease spread-RELAT mosquito aedis edippai COP.
<Translation> Disease spreading mosquito is Aedis Edippai.
<SMT-ENG> Disease and mosquito aedes aegypti .
<NMT-ENG> Malaria spreads while mosquito infected person .

4.1.3 Coordination constructions

The co-ordination constructions at clausal level are consistently translated incorrectly in both the directions in all cases of the co-ordination sample set. The construction is realised in Malayalam with complex syntactic form. The particle suffix -um is attached to all coordinating elements, but the same particle is used as an emphatic particle and also as an inclusion purpose as well. Apart from these usages the particle -um is also used for the future tense inflection. It might be the reason none of the usages of -um is translated correctly.
Example:
<SRC>ഓക്കാനം അഥവാ ഛർ ദിക്കാന് തോന്നും
<Gloss> Nausea or vomit-INF feel-FUT.
<Translation> Nausea or vomitting will feel.
<NMT-ENG> Nausea and vomiting .

4.1.4 Semantic handling in translation

A significant number of outputs generated by SMT and NMT depict correct syntactic structures but have a potent semantic loss. This is another important challenge, since the sentences being translated look like the correct usage in the target language, but the intended meaning has absolutely changed. NMT displays more such occurrences when compared to SMT, and often fails to realise the correct semantic role in the target language.

<SRC> ഇത്തരത്തിലുള്ള രോഗികളിൽ സമയത്ത് അന് റീബയോട്ടിക് ഔഷധം നല് കണം

<Gloss> This-kind-EXT-RELATE patients anti-biotic medicine give-IMP.

<Translation>This kind of patients anti-biotic medicines should be given at right time.

<SMT> Such രോഗികളിൽ time അന് റീബയോട്ടിക് medicine നല് treatment .

<Gloss> Such "patients-LOC" time "anti-biotic" medicine give-ROOT treatment .

<NMT-ENG> In such a case the medicine should be given to the doctor .

4.2 Hindi to Malayalam translation

The NMT performance on simple sentences and sentences with postpositional phrases are reasonably good, except few cases of complex syntactical co-ordination constructions and complex predicates.

Example-1:

<SRC> जिला मानसिक स्वास्थ्य कार्यक्रम के लिए उत्तराखंड के दो जिलों का चयन किया गया है।

<Translation> Two districts of Uttarakhand have been selected for District mental health programme.

<NMT> ജില്ലാ മാനസികാരോഗ്യ പരിപാടിയിൽ ഉത്തരാഖണ്ഡിലെ രണ്ട് ജില്ലകൾ തെരഞ്ഞെടുത്തു

<Gloss> District mental-health programme-LOC Uttarakhand two districts selected.

<Translation> In distirict mental-health programme two districts located in Uttarakhand selected.

<SMT> ജില്ലാ മാനസികാരോഗ്യ പരിപാടിക്കകത്ത് വേണ്ടി ഉത്തരാഖണ്ഡിലെ രണ്ടു ജില്ലകളിൽ തിരഞ്ഞടുക്കുന്നത് ചെയ്തു

<Gloss> District mental-health programme-within for Uttarakhand-LOC two districts-LOC select-NOMIN did.

<Translation> For within the district mental-health programme selecting two districts in Uttarakhand that did.

Example-2:

<SRC> टीके लेने की सलाह भी दी जाती है।

<NMT> വാക്സിൻ എടുക്കുന്നതിനുള്ള ഉപദേശം നൽകണം

<Translation> Advice should be given for taking vaccine

<Gloss> vaccine take-NOMIN-EXIST-REL advice give-IMP.

<SMT> കുത്തിവയ്പ് എടുക്കുന്നതിനുള്ള ഉപദേശം എന്നിവയും നൽകപ്പെടുന്നു

<Translation> Advice for taking Vaccination and so on are given.

<Gloss> Vaccination take-NOMIN-EXIST-RELAT advice so-on-also give-PASS-PRS.

Example-3:

<SRC> कुष्ठ रोग के लक्षणों को पहचानना बहुत कठिन है।

<NMT> കുഷ്ഠരോഗ ലക്ഷണങ്ങൾ തിരിച്ചറിയുന്നത് വളരെ അപൂർവ്വമാണ്

<Translation> Identifying leprosy syptoms that is very rare.

<Gloss> Leprosy symptoms identify-NOMIN very rare-COP.

<SMT> കുഷ്ഠ രോഗത്തിൻറെ ലക്ഷണങ്ങൾ തിരിച്ചറിയുന്നത് വളരെ കഠിനമാണ്

<Translation> Identifying leprosy's symptoms that is very hard.

<Gloss> Leprosy disease symptoms identify-NOMIN very dificult-COP.

4.2.1 Co-ordination constructions

Hindi sentences with co-ordination constructions are incorrectly translated to equivalent Malayalam sentences. This is in alignment with the previous observation that complex syntactic sentence constructions are mostly translated incorrectly. This might be due to the vast differences in the way co-ordination constructions are realised in two languages.

<SRC> माटा एनाफिलिक मच्छर स्वच्छ पानी में पैदा होता है और रात में काटता है।

<NMT> മലാരിയില് നിന്ന് വല ശേഖരിക്കരുത്

<Translation> From Malaria do not collect net.

<Gloss> Malaria-LOC from net collect-NEG.
<SMT> കൊതുക് വൃത്തിയുള്ള വെള്ളത്തിൽ ജനിക്കുകയും രാത്രിയിൽ എന്നിവ ഉണ്ടാകുന്നു
<Translation> Mosquito takes births also in fresh water night and so on makes.
<Gloss> Mosquito fresh-EXIST-RELAT water-LOC take-birth-COORD night-LOC so-on make-INTR-PRS.
<Gloss> Mosquito clean-EXIST-RELAT water-LOC take-birth-CORD night-LOC and-so-on make-INTR-PRS.

4.2.2 Complex predicate constructions

Hindi complex predicate constructions are the other set of constructions which consistently fail to be translated correctly from Hindi to Malayalam. The usage of complex predicate is confined to Hindi and the equivalent Malayalam is obtain by relative clause. The different ways of realising the semantically equal sentence might lead to problems. Secondly, scrambling of arguments with in a clause is usual in relatively free word order languages. It may also lead to the failure of this translation. The erroneous outputs are shown in the example-1 and example-2.
Example-1
<SRC> कुष्ठ रोग एक प्रकार के वायरस से होने वाला एक रोग है।
<NMT> വൈറസ് ഒരു തരത്തിലുള്ള വൈറസ് വൈറസ് ആണ്
<Translation> Virus one kind of virus is virus.
<Gloss> Virus one kind-EXIST-RELAT virus virus COP.
<SMT> കുഷ്ഠരോഗം ഒരു തരത്തിലുള്ള വൈറസ് കൊണ്ട് ഉണ്ടാകുന്ന ഒരു രോഗമാണ്
<Translation> Leprosy is due to one kind of virus making disease.
<Gloss> Leprosy one kind-EXIST-RELAT virus by make-INTR-PRS one disease-COP.
Example-2
<SRC> कुष्ठ रोग वायरस से होने वाला एक प्रकार का रोगजनन है।
<NMT> വൈറസ് വൈറസ് ബാധിച്ച ഒരു തരത്തിലുള്ള രോഗമാണ്
<Translation> Virus virus effected one kind disease is.
<Gloss> Virus virus infect-RELAT one kind-EXIST-RELAT disease-COP.
<SMT> കുഷ്ഠരോഗം വൈറസ് മൂലമുണ്ടാകുന്ന ഒരു തരത്തിലുള്ള
<Translation> Leprosy due to virus one kind of.
<Gloss> Leprosy virus due-to-make-INTR-PRS-RELAT one kind-EXIST-RELAT.

4.3 Malayalam to Hindi translation

The translations from Malayalam to Hindi using NMT do not perform better than Hindi to Malayalam. Verbal inflections are in Malayalam are not able to be translated in the apt manner in Hindi. However, certain simple sentences are handled reasonably well by NMT.
Example-1:
<SRC> നിങ്ങള് കമ്പിളി വസ്ത്രങ്ങള് ധരിക്കുക
<Gloss> you woolen clothes wear-IMP.
<Translation> You may wear woolen clothes.
<NMT-HND> आप ऊनी कपड़े पहनें ।

Example-2:
<SRC> ഉറങ്ങുന്ന സമയത്ത് കൊതുകു വല ഉപയോഗികണ
<Gloss> Sleep-RELAT time-DAT mosquito net use-IMP.
<Translation> while sleeping mosquito net should be used.
<NMT-HND> सोते समय मच्छरदानी पम्प करें ।

Example-3:
«Semantic error»>
<SRC> ഇത്തരത്തിലുള്ള രോഗികളിൽ സമയത്ത് അന് റീബയോട്ടിക് ഔഷധം നല്കണ
<Translation> This-kind-EXIST-RELAT disease-LOC time-DAT antibiotic medicine give-IMP.
<NMT-HND> ऐसे सहवास के समय एंटीबायोटिक्स औषधि देनी चाहिए ।

4.3.1 Coordination constructions

Simple nominal co-ordination constructions are successfully translated from Malayalam to Hindi. However, the complex sentential coordination is still out of its reach. The example-1 nominal coordination is translated well, whereas it failes on example-2.

Example-1:
<SRC> ഓക്കാനം അഥവാ ഛർദ്ദിക്കാന് തോന്നും
<Gloss> Nausea or vomit-INF feel-FUT.
<Translation> You will feel Nausea or vomitting.
<NMT-HND> मितली या उल्टी महसूस होना ।
Example-2:
<SRC> ഓക്കാനവും ഛർദ്ദിയും ഉണ്ടാകു
<Translation> It makes nausea or vomitting.
<NMT-HND> मतली की नियुक्ति

4.3.2 Cleft constructions

Cleft sentences in Malayalam are still a problem for Malayalam to Hindi translation. The simple sentence is translated correctly (Example 1), but its cleft form is translated incorrectly in (Example 2). This shows the dire need of an approach which can enhance NMT to learn different syntactical constructs prevalent in linguistically distant languages.
<SRC> അവന് ഡങ്കി പനി ബാധിച്ചു
<Gloss> He-DAT Dengue fever caught.
<Translation> He caught dengue fever.
<NMT-HND> जिसके कारण डेंगू बढ़ता है।
<SRC> അവനാണ് ഡങ്കി പനി ബാധിച്ചത്
<Gloss> He-COP dengue fever caught-NOMIN.
<Translation> It is him that caught fever.
<NMT-HND> वह डेंगू बुखार है।

5 Conclusion

Based on the extensive evaluation carried out on the NMT bi-directional translator with possible pairs of English, Hindi, and Malayalam, simple sentences including sentences with complex postpositional phrases are translated well in all pairs. The output quality is consistently better than SMT in most of the phenomena. An exceptional case is shown in the modal affixes of Malayalam, which are translated incorrectly to Hindi. The other important observation is that NMT is not able to decode complex verbal inflections and translate them to the target language, particularly to Hindi. A major issue of NMT is that it can not translate complex syntactic structures, particular to the source language usage. It is visible from the cleft and participle constructions of Malayalam failing to get translated to other languages and similarly complex predicate structures in Hindi to other languages. In addition to these, co-ordination constructions with conjuncts are also translated incorrectly by both SMT and NMT. These factors can serve as important guidelines to be considered when building parallel corpora for linguistically distant languages in the future, to facilitate better performance of SMT as well as NMT approaches on these language pairs.

References

Raghavachari Amritavalli. 2014. Separating tense and finiteness: anchoring in dravidian. *Natural Language & Linguistic Theory*, 32(1):283–306.

R Amritavalli. 2017. 9 nominal and interrogative complements in. *Dravidian Syntax and Universal Grammar*.

Ronald E Asher and TC Kumari. 1997. *Malayalam*. Psychology Press.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1994. Anusaraka or language accessor: A short introduction. *Automatic Translation, Thiruvananthpuram, Int. school of Dravidian Linguistics*.

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal, and KV Ramakrishnamacharyulu. 1995. *Natural language processing: a Paninian perspective*. Prentice-Hall of India New Delhi.

Junyoung Chung, Caglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *ICML*, pages 2067–2075.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1152–1161. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics.

KA Jayaseelan and R Amritavalli. 2005. Scrambling in the cleft construction in dravidian. *The free word order phenomenon: Its syntactic sources and diversity*, 69:137.

Karattuparambil A Jayaseelan. 2001. Ip-internal topic and focus phrases. *Studia Linguistica*, 55(1):39–75.

KA Jayaseelan. 2014. Coordination, relativization and finiteness in dravidian. *Natural Language & Linguistic Theory*, 32(1):191–211.

Girish Nath Jha. 2010. The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Sata-anuvadak: Tackling multiway translation of indian languages. *pan*, 841(54,570):4–135.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Thomas McFadden and Sandhya Sundaresan. 2014. Finiteness in south asian languages: an introduction. *Natural Language & Linguistic Theory*, 32(1):1–27.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Interspeech*, pages 194–197.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Judith Tonhauser. 2015. Cross-linguistic temporal reference. *linguistics*, 1(1):129–154.

Paul J Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78:1550–1560.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.