Translation Divergence in English-Sanskrit-Hindi language pairs

Pawan Goyal[1] and R. Mahesh K. Sinha[2]

[1] School of Computing and Intelligent Systems, University of Ulster, UK,
goyal-p@email.ulster.ac.uk
[2] Indian Institute of Technology, Kanpur, India,
rmk@iitk.ac.in

Abstract. The development of a machine translation system needs that we identify the patterns of divergence between two languages. Though a number of MT developers have given attention to this problem, it is difficult to derive general strategies which can be used for any language pair. Therefore, further exploration is always needed to identify different sources of translation divergence in different pairs of translation languages. In this paper, we discuss translation pattern between English-Sanskrit and Hindi-Sanskrit of various constructions to identify the divergence in English-Sanskrit-Hindi language pairs. This will enable us to come up with strategies to handle these situations and coming up with correct translation. The base has been the classification of translation divergence presented by Dorr [Dorr, 1994].
keywords: Machine Translation, Translation Divergence

1 Introduction

Translation divergence occurs when the underlying concept of a sentence gets manifested differently in different languages. The topic has been studied from different perspectives and a number of approaches have been proposed to handle them [Habash and Dorr, 2002]. It is difficult to obtain correct machine translation for any MT system without identifying the nature of translation divergence. In this paper, we examine English-Sanskrit and Hindi-Sanskrit language pairs mostly from the perspective of identifying the language specific divergences. The languages, English and Sanskrit, as well as Hindi and Sanskrit differ in many respects, presenting a rich source for the study of translation divergence in MT. In section 2, we look at the translation divergence classification proposed by Dorr and in what way, it appears in the English-Sanskrit-Hindi language pairs. In section 3, we look at some other divergence patterns which cannot be classified in the divergence patterns identified by Dorr. In section 4, we give the concluding remarks.

2 Divergence patterns identified by Dorr

Dorr has identified seven classes of translation divergences. These classes are:

1. Thematic Divergence
2. Promotional Divergence
3. Demotional Divergence
4. Structural Divergence
5. Conflational Divergence
6. Categorial Divergence
7. Lexical Divergence

These classes have been defined to account for different types of translation divergences found in a pair of translation languages. Let us look, in what sense these can be seen while we examine English-Sanskrit and Hindi-Sanskrit language pairs.

2.1 Thematic Divergence

Thematic divergence refers to the divergences arising from differences in the realization of the argument structure of a verb. In the language pairs, we have considered, we can find many examples of this divergence. Let us consider some of these:

– Subject NP in English in nominative case while subject NP in Sanskrit in dative case:

1. I like sweets. ⇒ *mahyam madhuram rocate.*
   (I DAT) (sweets) (like.PR[1]).
   ⇒ Sweets are liked by me.

When we go from Hindi to Sanskrit, same divergence appears:

2. *main miṭhāī̄ pasanda karatā hūM.* ⇒*mahyam madhuram rocate.*
   (I) (sweets) (like) (do) (be.PR) (I DAT) (sweets) (like.PR).
   ⇒ *mujhe miṭhāī̄ pasanda haiṃ.*
   (I DAT) (sweets) (like.PR).

From the example 2, we see that there is divergence between Hindi and Sanskrit. Here the experiencial verb '*ruc*' gets an active construction in Hindi while it conditions a dative subject in Sanskrit. However, there is no divergence when we go from Sanskrit to Hindi since the closest translation in Hindi is *mujhe miṭhāī̄ pasanda haiṃ* which has a dative subject as well.
If we examine other verbs such as 'eat' for the same pattern, we will not find this divergence.

3. I eat sweets. ⇒*aham madhuram khādāmi.*
   (I) (sweets) (eat.PR)
   ⇒ *main miṭhāī̄ khātā hūM*
   (I) (sweets) (eat.PR)

Thus, this divergence is present in a special category of verbs and not all the verbs. *Pāṇini* in his *kāraka adhikāra* lists special cases of verbs which require special treatment. In a work on English verb classes by Levin [Levin, 1997], semantic classes of verbs are analyzed which give rise to Divergence.

2.2 Structural divergence

These are the examples where a noun phrase (NP) is realized in different ways in two languages. This is most common between English and Sanskrit because in Sanskrit, no noun is pronounced without a *vibhakti*. This *vibhakti* can be realized in English either as a null or a preposition. Here are some of the examples that exhibit Structural divergence:

4. He brought mangoes.	⇔ *saḥ āmrāni ānayat.* (He) (mangoes) (bring.PST)
5. He went to the market.	⇔ *saḥ āpaṇam agacchat.* (He) (market.ACC) (go.PST)
6. He enters the class.	⇔ *saḥ kakṣāyām praviśati.* (He) (class.LOC) (enter.PR)

In example 4 and 6, the *vibhakti* in Sanskrit is not realized by a preposition in English but a null, while in example 5, the *vibhakti* is realized by a preposition in English.
While we go from Sanskrit to hindi, we will not get many examples of this divergence since both languages are *kāraka* and *vibhakti* based.

2.3 Conflational and Inflational Divergence

A conflational divergence results when two or more words in English are translated by one word in Sanskrit. There are many mechanisms in Sanskrit that present this divergence. Same is true between Hindi and Sanskrit too. Let us look at some of them:

– '***sannata prayoga***':-
7. *aham pipaṭhiṣāmi* ⇒ I want to read.
(I) (want to read)

⇒ aham paṭhitum icchāmi
(I) (to read) (want)

aham pipaṭhiṣāmi ⇒ *main paḍhanā cāhatā hūM.*
(I) (want to read) (I) (read) (want be.PR)

⇒ aham paṭhitum icchāmi
(I) (to read) (want)

[1] Appendix 1

A sentence in Sanskrit such as '*aham pipaṭhiṣāmi*' is translated in English as 'I want to read', thus 'want to see' is translated as '*pipaṭhiṣāmi*'. Thus it presents conflational divergence. This divergence is exhibited even between Hindi and Sanskrit.

– '***nāmadhātu prakriyā***':- In certain meanings, a Sanskrit nominal stem can accept certain *pratyayas*. Consider the sentences:

8. *saḥ paṇḍitāyate* ⇒ He behaves like a scholar.
(He) (behaves like a scholar)
⇒ *saḥ paṇḍitāḥ iva ācarati*
(He) (Pandita) (like) (behave.PR)

saḥ paṇḍitāyate ⇒ *vaha paṇḍita kī taraha ācaraṇa karatā hai.*
(He) (behaves like a scholar) (He) (scholar) (like) (behave do be.PR)
⇒ *saḥ paṇḍitāḥ iva ācarati*
(He) (Pandita) (like) (behave.PR)

He behaves like a scholar.⇐ *saḥ paṇḍitāyate*. The affix *kyaṅ* is applied to the noun *paṇḍita* in the sense of 'behaves like'.

9. *saḥ śiṣyam putrīyati* ⇒ He treats the disciple as his son.
(He) (disciple.ACC) (treats as son)
⇒ *saḥ śiṣyam putram iva ācarati*
(He) (disciple.ACC) (son) (like) (behave.PR)

saḥ śiṣyam putrīyati ⇒ He treats the disciple as his son.
The affix *kyac* is applied to the noun *putra* in the sense of 'treats like'.

In example 9, 'treats as son' in English is realized by a single word '*putrīyati*' in Sanskrit and presents an example of conflational divergence.

– '***yaṅanta prayoga***':- This refers to frequentatives.

10. *saḥ pāpacyate* ⇒he cooks again and again
(He) (cooks again and again)
⇒ *saḥ punaḥ punaḥ pacati*
(He) (again) (again) (cook.PR)

Same pattern is exhibited when we go from Sanskrit to Hindi language pair.
saḥ pāpacyate ⇒ *vaha bāra bāra pakātā hai*
(He) (cooks again and again) (He) (again) (again) (cook be.PR)
⇒ *saḥ punaḥ punaḥ pacati*
(He) (again) (again) (cook.PR)

Hence the English words, 'cooks again and again' and the Hindi words, '*bāra bāra pakātā hai*' will be translated in Sanskrit as '*pāpacyate*'. Here, the root 'pac (to cook)' is applied with the affix '*yaṅ*' to form '*pāpacyate*'. This again,

exhibits the example of conflational divergence.

2.4 Categorial Divergence

Categorial divergences are located in the mismatch between parts of speech of the pair of translation languages. Consider the following example:

11. She is jealous of me. ⇔ *sā mahyam īrṣyati*
(She) (with me) (jealousy does).

We notice that in Sanskrit, 'jealous' is realized by a verbal mapping, thus presenting categorial divergence. When we go from Hindi to Sanskrit, we have another translation possible in hindi, for example:

usako mujhase īrṣyā hai ⇒ *sā mahyam īrṣyati*
(She.DAT) (me-from) (jealousy be.PR) (She) (with me) (jealousy does).
⇒ *vaha mujhase īrṣyā karatī hai*
(she) (me-with) (jealousy) (do)

2.5 Lexical Divergence

Lexical divergence arises out of the unavailability of an exact translation map for a construction in one language into another language. In Sanskrit, by adding '*upasarga*' to a verb, it gets a different meaning. For example, consider the following sentences:

12. *saḥ mām vadati* ⇒ He speaks to me.
(He) (me.ACC) (speaks)

13. *saḥ kṣetre vivadati* ⇒ He quarrels in the field.
(He) (field.LOC) (quarrel.PR)
⇒ *saḥ kṣetre kalahaṃ karoti*
(He) (field.LOC) (quarrel) (do)

In example 12, the Sanskrit verb '*vad*' is realized by English verb 'speak', while in example 13, the Sanskrit verb '*vi+vad*' (*upasarga 'vi'* is added to verb *vad*) is realized by a new verb in English 'quarrel'.

3 Other Divergence Patterns

Let us look at some other divergence patterns that are found in these languages:

3.1 Implications of word order

Though Sanskrit is a free phrase order language, there are situations where the word order changes the meaning of the sentence. An example is the occurrence

of *kim*, consider the sentences:

14. *kim saḥ khādati?* ⇔ 'is he eating?'
(QP) (he) (eat.PR)

15. *saḥ kim khādati?* ⇔ 'what is he eating?'
(He) (IP) (eat CONT)

16. *saḥ khādati kim?* ⇔ 'is he eating?'
(He) (eats) (QP)

Thus, two different interrogative patterns in English are taken care of by different word orders. Similar situation does not arise when we examine Hindi-Sanskrit language pair and we donot find this divergence. Thus, we have:

14. *kim saḥ khādati?* ⇔ '*kyā vaha khā rahā hai?*'
(QP) (he) (eat.PR) (QP) (he) (eat) (PROG) (be.PR)

15. *saḥ kim khādati?* ⇔ *vaha kyā khā rahā hai?*
(He) (IP) (eat CONT) (He) (IP) (eat)(PROG) (be.PR)

16. *saḥ khādati kim?* ⇔ *vaha khā rahā hai kyā?*
(He) (eats) (QP) (He) (IP) (eat)(PROG) (be.PR) (IP)

In Sanskrit, word order is used to decide for definiteness for a noun. For example, consider the two sentences:

17. *bālakaḥ gṛhe asti* ⇔ 'The boy is in the house'
(boy) (house.LOC) (be.PR)

18. *gṛhe bālakaḥ asti* ⇒ 'A boy is in the house'.
(house.LOC) (boy) (be.PR)

Thus, *bālakaḥ* occurs at different positions in the sentence to show 'a definite boy' (example 17) and 'some boy' (example 18). In other words, the bare noun phrase '*bālaka*' in 17 and 18 is mapped by definite and indefinite noun phrases in English. However, the only difference between these two Sanskrit sentences is the respective positions of the subject NP and the adverbial phrase. When we look at the reverse translation of 18, we find that the nature of divergence is different. Thus, we have:

19. A boy is in the house. ⇒ *gṛhe ekaḥ bālakaḥ asti*
(house.LOC) (a) (boy) (be.PR)

On the other hand, there is no divergence between Hindi-Sanskrit language pair on this issue.

3.2 Change of voice

In Sanskrit language, we find the use of passive voice to be very frequent, which is not so in English. We are presenting the examples below which show divergence when we go from Sanskrit to English translation. The Sanskrit sentence is in passive voice, while the corresponding sentence in English sentence is in active voice.

20a. *rāmeṇa hasitavyam* ⇒ Ram should smile.
(Ram.INS)(laugh.KR)

⇒*rāmaḥ haset.*
(Ram) (smile.IMPR)

21a. *kopaḥ na karaṇīyaḥ bhavatā* ⇒ You should not be angry.
(anger) (not) (do.KR) (you.INS)

⇒ *tvam mā krudhya*
(you) (not) (anger.IMPR)

22a. *tena khāditaḥ* ⇒ He ate.
(he.INS) (eat.PASS)

⇒ *saḥ akhādat.*
(he) (eat.PST)

While examining Hindi-Sanskrit language pair, we do not find similar divergence since the corresponding Hindi sentences are very close to the passive construct in Sanskrit:

20b. *rāmeṇa hasitavyam* ⇒ *rāma ko haṃsanā chāhie*
(Ram.INS)(laugh.KR) (Ram.ACC) (laugh) (should)

21b. *kopaḥ na karaṇīyaḥ bhavatā* ⇒ *āpako gussā nahīṃ karanā chāhie*
(anger) (not) (do.KR) (you.INS) (you.ACC) (anger) (not) (do) (should)

22b. *tena khāditaḥ* ⇒ *usane khāyā*
(he.INS) (eat.PASS) (He) (eat.PST)

3.3 Gerunds and Participle Clauses

Another significant source of divergence in Sanskrit and English/Hindi can be located in the way various clauses and adjuncts are realized in different languages. First, let us consider English and Sanskrit language pair:

23. 'He is happy **to protect** the country'
⇔ *deśam* ***rakṣitvā*** *saḥ prasannaḥ bhaviṣyati.*
(country.ACC) (protect.GER) (he) (happy) (be.FU)

24. 'He came here **to protect** the country'
⇔ *deśam* ***rakṣitum*** *saḥ atra āgacchat*
(country.ACC) (protect.GER) (he) (here) (come.PST)

25a. He is not able **to walk**.
⇔ *saḥ* ***calitum*** *asamarthaḥ.*
(He) (walk.GER) (not able)

We notice that in example 23 and 24, in Sanskrit, different types of adjunct verbal clauses and complement verbal clauses are realized by different structures. In English, they are realized by an infinitive clause. The examples 24 and 25a have similar sentence construction. We now examine some sentences between Sanskrit and Hindi languages:

25b. *vaha* ***calane meṃ*** *asamartha hai* ⇔ *saḥ* ***calitum*** *asamarthaḥ.*
(He) (walk) (in) (able) (not) (be.PR) (He) (walk.GER) (not able)

26. *vaha citra* ***dekhane (ke liye)*** *āyā* ⇔ *saḥ citram* ***draṣṭum*** *āgataḥ.*
(He) (picture) (see) (for) (come.PST) He) (picture) (see.GER) (come.PST)

In the Hindi sentences in (25b-26), the adjunct verbal clauses and complement verbal clauses are realized by different structures, which in Sanskrit are mapped by a single structure. Though, for example 25b, we have a Sanskrit parallel as '*saḥ calane akuśalaḥ*' which does not present divergence.

3.4 Morphological Gaps

We take the example of causatives:

27. 'I study' ⇒ *aham paṭhāmi*
(I) (study.PR)

28. 'I make him study' ⇒ *aham tam pāṭhayāmi*
(I) (He.ACC) (teach.PR)

In the above two sentences, the form *paṭhāmi* and *pāṭhayāmi* are morphologically derived from the root *paṭh*, while the English counterpart has only one lexical verb 'study' and other is derived using the verbs such as 'get', 'make' etc, with separate argument structure. In case of Hindi-Sanskrit, no divergence is exhibited as such since in Hindi also, roots are morphologically derived:
paḍhā ⇒ *paḍhāyā*⇒*paḍhavāyā.*

3.5 Honorific

In Sanskrit, honorific features are expressed by the use of plural pronoun (as well as adjective and noun, this is crucial since the verb endings need to agree with noun) and plural verb inflections. For example, consider the sentence:

29.Respected teacher is teaching the students.
⇔ *pūjyāḥ gurucaraṇāḥ śiṣyān pāṭhayanti.*
(respected.pl) (teacher.pl) (students.ACC) (teach.PR)
⇔ *pūjya gurujī ̄ ̀sìṣyoṃ ko paḍhāte haiṃ*
(respected) (teacher) (students) (to) (teach)

We see that in example 29, the adverb '*pūjya*', noun '*gurucaraṇa*' as well as the verb '*pāṭh*' take plural inflections in case of Sanskrit, while in hindi only the verb '*paḍhāte haiṃ*' takes the plural inflection. This divergence is caused by the socio-cultural aspect of the respective languages.

3.6 Mapping of Time

In English, the concept of a.m. vs p.m cannot be exactly mapped in Sanskrit. The example 30 shows that the time at 5 o'clock in the morning (*prātaḥkāle pañcavādane*) is denoted by a.m. in English. In example 31, the time at 11 o'clock in the morning/afternoon (*prātaḥkāle/madhyadine ekādaśavādane*) is also denoted by a.m. in English. Therefore, the term a.m. (and similarly p.m.) cannot be translated as such. One needs to examine the numbers written before and should have a built in intelligence in the translation system to handle different numbers by appropriate Sanskrit words.
When we go from Sanskrit to English translation, this divergence pattern is not exhibited since English also has more terms for periods of day than a.m. and p.m.

30. He arrived at 5 a.m. ⇒ *saḥ prātaḥkāle pañcavādane āgataḥ*
(He) (morning.LOC) (at 5 o'clock) (arrive.PST).
⇒ He came at 5 o'clock in the morning.

31. He arrived at 11 a.m. ⇔ *saḥ prātaḥkāle/madhyadine ekādaśavādane āgataḥ.*
(He) (morning/afternoon.LOC) (at 11 o'clock)
(arrive.PST).

A similar situation is seen with respect to the mapping of p.m. in the examples 32-34.

32.He arrived at 3 p.m. ⇔ *saḥ aparāhne trivādane āgataḥ.*
(He) (afternoon.LOC) (at 3 o'clock) (arrive.PST).

33. He arrived at 5 p.m. ⇒ *saḥ sāyaṃkāle pañcavādane āgataḥ*
(He) (evening.LOC) (at 5 o'clock) (arrive.PST).
⇒ He came at 5 o'clock in the evening.

34. He arrived at 11 p.m. ⇒ *saḥ rātrau ekādaśavādane āgataḥ*
(He) (night.LOC) (at 11 o'clock) (arrive.PST).
⇒ He came at 11 o'clock in the night.

However, there is no divergence in case of this mapping, when we examine Hindi-Sanskrit language pair.

4 Conclusions and Discussions

Above mentioned are some of the divergence patterns that we were able to classify. We have kept in view the classification of translation divergence proposed by Dorr and some of the works on Hindi-English divergence [Sinha and Thakur, 2005] [Dave et. al., 2002]. We are in the process of identifying other such patterns. These divergence patterns will be useful in our implementation of machine translation system from English to Sanskrit language. Some of the divergence study has been useful in the current implementation of our machine translation system [Goyal and Sinha, 2008] from English to Sanskrit.

References

[Dorr, 1994] Dorr, B.: Classification of Machine Translation Divergences and a Proposed Solution Computational Linguistics **20 (4)** (1994) 597–633.

[Habash and Dorr, 2002] Habash, N. and Dorr, B.: Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation Technical Report, LAMP 88 (2002).

[Dave et. al., 2002] Dave, S. and Parikh, J. and Bhattacharya, P.: Interlingua Based English-Hindi Machine Translation and Language Divergence, Journal of Machine Translation (JMT), **17** (2002).

[Levin, 1997] Levin, B.: English Verb Classes and Alterations : A Preliminary Investigation, The MIT Press (1997).

[Sinha and Thakur, 2005] Sinha, RMK and Thakur, A.: Translation Divergence in English-Hindi MT EAMT, Budapest, Hungary (2005).

[Goyal and Sinha, 2008] Goyal, P. and Sinha, RMK: A Study towards English to Sanskrit Machine Translation system. SISSCL (2008).

Appendix 1

ACC: Accusative Case, **INS**: Instrumental Case, **LOC**: Locative Case,**AFF**: Affirmative Case, **CAUS**: Causative Case, **CONT**: Continuative Aspect, **CPP**: Conjunctive Participal Particle, **ET:** Determiner, **DUR**: Durative Aspect; **EW**: Echo Word, **FU**: Future Tense,**DAT**: Dative Case, **DIT**: ditransitive Case, **ERG**: Ergative Case, **GER**: Gerund, **HAB**: Habitual Aspect, **IMP**: Imperfective Aspect, **IMPR**:.Imperative Mood, **PASS**: Passive Particle, **PR**: Present Case, **INT**: Interrogative, **OPT**: Optative Mood, **QP:** Question Particle, **RP**: Relative Pronoun, **SUBJ**: Subjunctive Mood, **TRS**: Transitive, **VPRT**:Verbal Participle, **KR:** *kṛtya pratyayānta* in Sanskrit