Applying Sanskrit Concepts for Reordering in MT

Akshar Bharati, Sukhada, Prajna Jha, Soma Paul and Dipti M Sharma
Language Technology Research Center
International Institute of Information Technology
Hyderabad, India
sukhada@research.iiit.ac.in, pragya.jha.jk@gmail.com
soma@iiit.ac.in, dipti@iiit.ac.in

Abstract

This paper presents a rule-based reordering approach for English-Hindi machine translation. We have used the concept of *pada,* from Pāṇinian Grammar to frame the reordering rules. A *pada* is a word form which is ready to participate in a sentence. The rules are generic enough to apply on any English-Indian language pair. We tested the rules on English-Hindi language pair and obtained better comprehensibility score as compared to Google Translate on the same test set. In assessing the effectiveness of the rules on padas which are analogous to minimal phrases in English, we achieved upto 93% accuracy on the test data.

1 Introduction

Like many other languages, Hindi word order differs from English. That is why in machine translation (MT), in addition to finding appropriate word senses, reordering of words plays a crucial role. Reordering in rule-based MT is a stage where the positions of the words of the source language (SL) sentence are reordered according to the target language (TL) syntax.

Hindi is a relatively free word-order language. The relations between constituents are marked explicitly with the help of postpositions. The morphological richness of the language allows the constituents to change their positions freely within a clause (Kachru, 2006; Chatterjee et al., 2007). In spite of that, not all the words can occur freely at any position (Kachru, 2006). For example,

- Postpositions follow their objects
- Verbs always precede their auxiliaries
- Modifiers such as prepositional modifiers, adjectival modifiers or determiners precede their head

For such reasons and many more described in Section 4, one needs to reorder the sentences to arrive at a natural translation.

Statistically trained systems give good results in less amount of time and manual effort. But any statistically trained system requires huge parallel corpus. Indian languages (ILs) lack a reasonable parallel corpus size for English and IL pairs. Hence, we explore a rule based approach for reordering taking insights from Pāṇinian Grammar (PG).

The available reordering approaches are discussed in Section 2. Our reordering approach is described in Section 3. Section 4 talks about major divergences between English and Hindi. Reordering rule formation is described in Section 5 and Section 6 presents experiments and results. Section 7 does error analysis and Section 8 concludes the paper.

2 Related Work

There are many approaches to handle TL word ordering. Some of them are described below:

1. Koehn et al (2003) perform reordering by using relative distortion probability distribution model trained from joint probability distribution model $\phi(\bar{e},\bar{f})$. Their model relies on the language model to produce words in right order (Koehn, 2009).

2. Kunchukuttan et al (2014) developed a phrase based system for English-Indian Language (henceforth En-IL) pairs that uses two extensions- (i) source reordering for En-IL translation using source side reordering rules developed by (Patel et al., 2013) and (ii) describing untranslated words for Indian-IL translation by using transliteration procedure.

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 357–366,
Trivandrum, India. December 2015.

3. A statistical translation model introduced by Yamada and Knight (2001) incorporates features based on syntax and converts SL parse tree according to TL with the help of probabilistic operations at required nodes using expectation-maximization algorithm. This model accepts parse trees as an input on which it performs child node reordering according to the TL.

4. Costa-Jussà and Fonollosa (2009) used an Ngram-based Reordering (NbR) method that uses SMT techniques to generate reordering graph, which utilizes word classes and NbR model for reordering. It produces an intermediate representation of source corpora where word-order of SL is represented more closely to the order of TL.

5. Universal Networking Language (UNL) is an interlingual representation of MT systems through semantic graphs, which comprise of semantic relations, attributes and universal words. The UNL approach transforms the lexicon into semantic hyper-graph which decides the word-order of the TL through parent-child positioning and relationship priority (Singh et al., 2007).

3 Our Approach

We present a rule based approach which is based on PG, particularly on the concept of *pada*. Though this paper presents an English-Hindi (En-Hi) reordering approach, we claim that the same system is generic enough to be used for any English-IL pair with some modifications.

PG analyses a word as a combination of a root/stem (*prakṛti*) and an affix (*pratyaya*) (Bharati et al., 2015). It is both the root and the affix that jointly denote the meaning – *prakŕtipratyayau sahārtham brūtaḥ* (Lit. root and affix together convey the meaning) (Rathore, 1998). PG deals with morphology which is not separated but is interlinked with syntax and semantics (Subrahmanyam, 1999). This in a way helps capture "how languages encode information" and "how the information flows in language". Let us explore it through example 1, taken from Bharati et al. (2015).

(1) a. mārjārāḥ mūṣakān
cat.PL,NOM rat.PL,ACC
mārayanti
kill.3,PL,PR
'Cats kill rats.'

b. mūṣakān mārjārāḥ mārayanti

c. mārjārāḥ mārayanti mūṣakān

d. mārayanti mārjārāḥ mūṣakān

e. mūṣakān mārayanti mārjārāḥ

f. mārayanti mūṣakān mārjārāḥ

The words in Sanskrit, in example 1a, can be ordered in any possible combination as shown through 1b to 1f. Still, they convey more or less the same meaning, but in English, changing the order of the subject and object as in "Rats kill cats" changes the meaning of the sentence all together.

The words in 1a have explicit morphemes called *vibhaktis* attached to them which mark the desired information explicitly. In other words, all the words have formed *padas* (described in Section 3.1), hence, they can occur freely at any position in the sentence. This gives us the clue to identity the relational morphemes and attach them with their relata.

3.1 *Pada* Formation

In PG, a *śabda* denotes a linguistic expression ranging from individual speech sound to an utterance (Singh, 1991) whereas a *pada* is a primary structural unit that occurs in actual sentences. Sanskrit has a grammatical rule *apadaṁ na prayuñjīta* (Dvivedi, 1953) which says: "a word which is not a *pada* should not be used in a sentence". Pāṇini defines a *pada* as follows: "A finished word form which is ready to participate in a sentence" (Mahavir, 1984; Bharati et al., 2015).

The sūtra *suptiṅantaṁ padam* (A 1.4.14i) states: "a word ending in *sup* (nominal suffix) or *tiṅ* (verbal suffix) is called a *pada*". The *sup* and *tiṅ* are the nominal and verbal grammatical inflections. The *sup* is the nominal case and the *tiṅ* is finite verb inflection. *Vibhakti* is a general term in Sanskrit used for both the nominal as well as verbal suffixes.

According to the sūtra (A. 1.4.14),

1. *Prātipadika* + *sup* = *subanta pada*
(nominal stem + nominal suffix = nominal *pada*)

2. *Dhātu* + *tiṅ* = *tiṅanta pada*
(verbal stem + finite verbal suffix = verbal *pada*)

Hence *pada* is a grammatically inflected word form which is ready to participate in a sentence expressing its role and relations with other words in the sentence.

3.2 The *subanta* and *tiṅanta pada*s in English

Let us take some English sentences and use Pāṇinian primitives such as *sup*, *tiṅ* and *pada* etc. to analyse them.

(2) Rama laughed.

The constituency tree diagram for sentence 2 is shown in Figure 1. The word *Rama* is the subject and thus does not carry any case marker. We are now looking for the criteria equivalent to the notion of *sup* Bharati et al (1996) show that English has the notion of generalized vibhakti. which corresponds to the *sup* suffixes in Sanskrit. The generalized vibhakti is realized either through subject[1] or object positions or through prepositions. Thus in sentence 2, *Rama* occurring at subject position seems to carry no *sup*, but according to Bharati et al., (1996), since it occurs at the subject position it carries a generalized vibhakti in terms of subject position, hence, it is a *subanta pada*.

The verb *laughed* has *-ed* suffix as a *tiṅ*, hence it can be called as a *tiṅanta* pada. See Figure 1 where each box represents an independent *pada*.

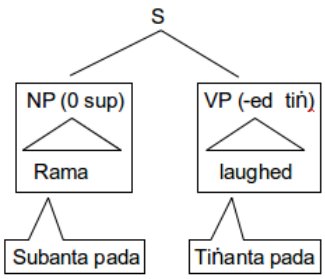

Figure 1: *Pada* information in tree diagram for 2

Example 2 contains only minimal phrases where each phrase consists of a single word, hence just by applying the sūtra *suptiṅantaṁ padam* (A 1.4.14) we can conclude that minimal phrase is a *pada* from Pāṇinian viewpoint. But what about the phrases which consist of more than one word? For example,

(3) The new students have been working on this problem.

[1]In linguistics, the notion of subject in ILs is much debatable (Bharati and Kulkarni, 2011).

In sentence 3, the word group *the new students* is an NP and *have been working on this problem* is a VP which consists of a verb group and a prepositional phrase. The verb group *have been working* contains *work* as a verb and *have been -ing* as a *tiṅ*/auxiliaries, hence, it can be taken as a *tiṅanta* pada. But, how many padas should we consider in the constituents *the new students* and *on this problem*?

The constituent *the new students* in sentence 3 occurs at subject position, hence, as stated by Bharati et al., (1996) it carries a generalized vibhakti in terms of subject position. Also, if the constituent *the new students* occurs in a prepositional phrase all three words take only a single preposition as in *I gave a book* ***to the new students***. The head noun *students* along with its modifiers *new* and *the* takes only one-vibhakti/*ekasup* 'to', Therefore, in example 3, the whole group, *the new students* can be taken as a single *subanta* pada. Similarly, *on this problem* can also be taken as a *subanta* pada which carries the preposition *on* as a *sup*.

As said before, a *tiṅanta pada* is formed by adding *tiṅ* suffixes to the verbal stems.

The finite verb phrases (VPs) are exceptional cases in English where an adverbial phrase can occur between a finite verb group. For example, take sentence 4:

(4) The child is impatiently
baccā *AUX* besabrī se
waiting for
pratīkṣā kara.3.SG.PR.CONT kī
her mother.
vaha.GEN.SG māṃ
'baccā besabrī se apanī māṃ kī pratīkṣā kara rahā hai'

In 4, the adverb 'impatiently' is embedded in the *tiṅanta pada*, *is waiting*. This is a frequent phenomenon in English. Hence, identification of *tiṅanta padas* helps in forming a verb group whereby translation of verb and its suffixes can be handled properly by grouping the verb root and verbal affixes the *tiṅ*.

In literature, a quite similar concept is mentioned by the name of Local Word Grouping (LWG), whereby word groups are formed based on the local (adjacent) word information (Bharati et al., 1995). In our attempts, we form word groups on the basis of nearby verbal and nominal suffixes called *sup* and *tiṅ* inspired by Pāṇinian

grammar which is linguistically more precise and satisfactory. This also facilitates the study of relational information that binds the words into a meaningful sentence.

In Hindi and Hindi like languages where the grammatical information is lexicalized, we group the postpositions with the nouns for *subanta* and auxiliaries with the verb for *tiṅanta*. Since auxiliaries and postpositions are grouped with the verbs and nouns, we get the padas and we can directly substitute them. In English also the auxiliaries carry the *tiṅ* information for the verbs. We are treating prepositions as the *sup* for nouns. In this way, the prepositions and postpositions for nouns and auxiliaries for verbs with respect to word order are taken care of by this step.

This approach handles grouping of noun with *sup* inflection and verb with *tiṅ* inflection. Hence, we do not need to have separate reordering rules for this. However, reordering between En-Hi pair is a more complex task, because, in English, a lot of relational information is encoded in position, which makes syntax of English very diverse from Hindi.

4 Major Divergences between English and Hindi

Language divergence includes lexical, structural and conflational divergences (Dorr, 1993). Dave et al (2001) discuss the major structural divergences with respect to English and Hindi. Table 1 summarizes the major structural divergences between English and Hindi.

English	**Hindi**
SVO	SOV
head first	head last
prepositional	postpositional
subject is sacrosanct	subject may be dropped

Table 1: Major structural divergences between English and Hindi

5 Formation of Reordering Rules

Hindi and English follow 'mirror structure' in terms of structural word order. Hence the daughters of English verb phrase (VP) and prepositional phrase (PP) come in mirror image when translated into Hindi. In other words, the SL VP looks like a mirror image of the TL VP and vice versa.

The concept of 'mirror structure' is illustrated with example 5. The constituency tree for sentence 5 is shown in Figure 2. The mirror image of its VP as shown in Figure 3 gives a perfectly ordered Hindi sentence.

(5) Queen Victoria opened Blackfriars Bridge in November 1869.

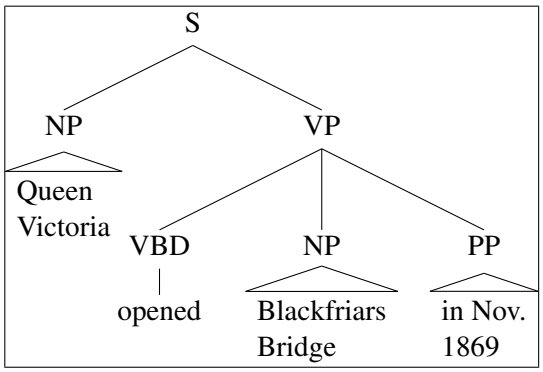

Figure 2: Showing constituency parse for 5

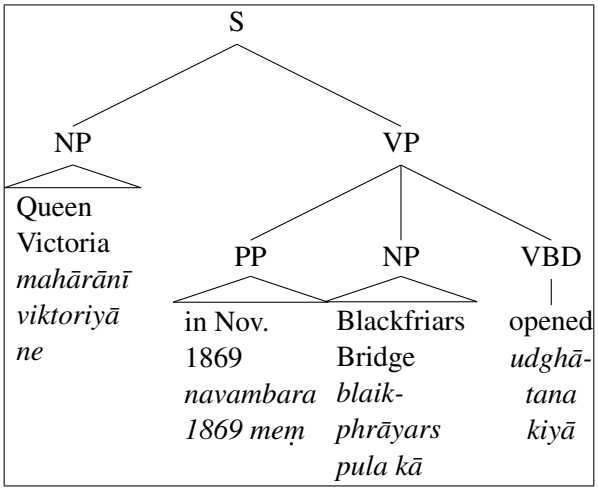

Figure 3: Showing mirror image of the constituency parse for 5

When mere 'mirror structure' does not suffice to arrive at a fluent Hindi sentence, we take advantage of dependency parse to handle such peculiarities in order to put the phrase at the desired place in TL. For example, indirect object tends to precede the direct object in Hindi. In 6a, the indirect and direct object reversal is restricted using dependency relations, since 6a sounds more natural than 6b to the native speakers of Hindi.

(6) a. Hari ne use eka
Hari.PROPN,NOM he.DAT a

billī dī
cat.ACC,SG give.PT,SG,FEM
'Hari gave her a cat.'

b. Hari ne eka billī
Hari.PROPN,NOM a cat.ACC,SG
use dī
he.DAT give.PT,SG,FEM
'Hari gave her a cat.'

5.1 Special Cases in Hindi

Apart from the cases described in Section 1, there are constructions which are not so flexible in terms of word order. This section describes some of them.

- Starting point and end point: While talking about a range or a span, the phrase describing an initial point precedes the phrase that describes the end point. This phenomenon seems to be common in all languages. Some such examples are shown from English in 7.

  (7) a. In Kashmir, fishing is a good business and the ideal season is **from April to October**.
  b. I shall continue to work **from 6 a.m. till midnight**, even if it kills me.
  c. **October to March** is the best time to visit the Jaipur city.
  d. In the south, Jammu is a transition zone **from the Indian plains to the Himalayas**.

  In 7, all the phrases shown in bold letters do not change their positions. Only forming the padas by identification of *sup* and *tiṅ* gives perfect word order in Hindi, hence, no movement is required.

- Relative order of arguments of a verb: The internal order of theme and recipient roles in English can be expressed in two ways: (i) through position as shown in 8 and (ii) through preposition as shown in 9.

  (8) John gave **Harry** a book.

  (9) John gave a book **to Harry**.

  In Hindi, 8 and 9 both the sentences can be translated as in 10 and 11 respectively. Since Hindi is morphologically rich, it marks relations through morphemes. Therefore, as mentioned above, the order can be relatively free.

  (10) John ne **Harry ko** eka kitāba dī.

  (11) John ne eka kitāba **Harry ko** dī.

  However, more natural one is 10. This is the default order in Hindi. This is consistent with the SOV order of Hindi.

- Source and destination phrases: Source and destination phrases tend to occur immediately before their verbal head as shown in 12 and 13. For instance, *vaha use skūla le gayā.* sounds more natural than *vaha skūla use le gayā.* in Hindi.

  (12) He **took** her **to the school**.
  vaha **le gayā** use **skūla**
  'vaha use **skūla le gayā**.'

  (13) He **picked** her **from the party**
  vaha **le āyā** use **se** **pārtī**
  'vaha use **pārtī se le āyā**.'

  Whereas if source and destination both occur as arguments of a same verb then destination phrase tends to occur immediately before the verb as shown in 14.

  (14) He **took** her **home** from the
  vaha **le gayā** use **ghara** se
  party.
  pārtī
  'vaha use pārtī se **ghara le gayā**.'

- Marking negation: In English negation markers 'no' and 'not' occur with nouns and verbs respectively. Whereas in Hindi, if the head of the negation marker 'no' is a verb modifier then the negation marker comes before the verb. There also, if the verb is a conjunct verb (Begum et al., 2011) then it come before the helping verb as shown in 17.

  (15) **No** politician is
  koi nahīṃ rājanetā hai
  completely honest.
  pūrī taraha se īmānadāra
  '**koi** rājanetā pūrī taraha se īmānadāra
  nahīṃ hai'

(16) We seek **no**
hama cāhate haiṃ **koi nahīṃ**
reward.
ināma
'hama **koi** ināma **nahīṃ** cāhate haiṃ'

(17) He did **not** **wait**
vaha *PAST* **nahīṃ pratīkṡā kara**
for me.
liye mere
'usane mere liye **pratīkṡā nahīṃ kī**'

- Yes/no interrogative question: The yes/no interrogative morpheme is missing in English (Anantpur, 2009). Therefore, it inverts the positions of subject and verb/auxiliary to mark a yes/no interrogative. Whereas in Hindi, yes/no interrogative question marker *kyā* is lexicalized and normally occurs at sentence initial position. As a consequence, in addition to word order, one needs to insert the morpheme *kyā* in Hindi, as shown in 18.

(18) Did you eat?
PAST,yes/no āpane khā
'**kyā** āpane khāyā?'

Section 5.2 and 5.3 describe some important reordering rules which are sufficient to give an overview of our approach.

5.2 Rules Based on 'Mirror Structure'

The rules based on 'mirror structure' are very productive. These rules give around 70% constituent reordering accuracy.

We followed Penn tagset to represent constituents. The rules are written using the following convention: *[.mother child1 child2 child3]* → *[.mother child3 child2 child1]*, where the part before '→' denotes SL parse tree and the later part denotes TL parse tree. The item marked with dot (.) represents the mother node followed by its child nodes. The rules are written in following format: first, we give the rule and then the rules are described with an example/s.

RBMS1[2]: VP is reversed to allow 'mirror structure' as discussed in Section 3. See example 5.

RBMS2: [.ADVP RB NP] → [.ADVP NP RB]
RB should be the last child of an ADVP.

(19) It was **down about** **35 points**.
yaha thā **nīce** **lagabhaga 35 aṃka**
'yaha **lagabhaga 35 aṃka nīce** thā.'

RBMS3: [.SBAR IN S] → [.SBAR S IN] iff *IN* != sentential conjunction

SBAR introduced by preposition is reversed as in 20, except when it is not a sentential conjunction as in 21.

(20) I expect a rough market **before**
maiṃ āṡā kara asthira bājāra **pahale**
prices stabilize.
mūlya sthira ho
'maiṃ **mūlya sthira hone se pahale** asthira bājāra ki āṡā karatā hūṃ'

(21) It was so dark **that** I could
thaā itanā andherā **ki** maiṃ saka
not see anything.
nahīṃ dekha.PT kucha
'itanā andherā thā **ki** maiṃ kucha nahīṃ dekha sakā'

RBMS4: [.NP [.NP *] [.PP *] [.* *]] → [.NP [.* *] [.PP *] [.NP *]]

(22) The boy **in** **blue shirt** is here.
ladakā **vālā nīlī** **kamīja** hai yahāṃ
'**nīlī kamīja vālā** ladakā yahāṃ hai

5.3 Rules for Handling Exceptional Cases

Even after applying 'mirror structure' on VP, some cases remain non-fluent and/or incomprehensible because of the inflexibility of some of the constituents to precede/follow the other constituents as pointed out in Section 1. This section describes some rules which are exceptions of 'mirror structure'.

RRMS1: [.VP * V] → [.VP * V]

Since Hindi is a verb final language, we restrict VP inversion iff verb is its last child, as in 23.

(23) The prices of winter wheat **now**
mūlya.PL kā ṡarada geṃhūṃ **aba**
being **planted** will not fall soon.
-jā rahe bo.PSSV *FUT* nahiṃ gira jaldī
'**aba boye jā rahe** ṡarada geṃhūṃ ke mūlya jaldī nahiṃ gireṃge'

RRMS2: [.ADJP RB JJ PP] → [.ADJP PP RB JJ]

ADJP having a PP clause, is reversed in order to emphasize PP or sentential clause. See 24.

[2]The rules based on 'mirror structure are named as *RBMS + a number* and the rules for restricting 'mirror structure' are named as *RRMS + a number.*

(24) She is **very good at her** vaha hai **bahuta acchī meṃ apane work**.
kārya
vaha **apne kārya meṃ bahuta acchī** hai

RRMS3: [.SQ WHNP VBD NP VP] → [.SQ VBD NP WHNP VP]

Wh-element of a sentence should be placed before the verb of the sentence/clause.

(25) **How many people** did you see?
kitane loga *PT* āpa mila
'āpa **kitane logoṃ** se mile?'

5.4 Illustration of Reordering Rules

The rules are written using CLIPS (Giarratano and Riley, 1998) where higher precedence rules are given higher salience. In general, specific rules have higher salience than a general rule. An overview of the reordering rules is given in Table 2 where first column shows the reordering rules or procedure and second column shows the effect of that particular step.

6 Experiments and Results

We picked 500 Full-text sentences from COCA (Davies, 2010) as SL input. These sentences were divided into 20 sets of 25 sentences each. All these input sets were translated in TL Hindi, using our system and Google Translate. One human evaluator evaluated two different sets and each set was evaluated by three evaluators. In total 30 evaluators rated the translation quality of both the systems on a scale of 0-4 (Bharati et al., 2004). As per Bharati et al (2004) , this scale assesses translation quality in terms of comprehensibility. Note that the identity of any system was not disclosed to the human evaluators for unbiased ratings. We found that on this scale Google Translate obtained 44.2% comprehensibility score while our system obtained 51.8% comprehensibility score.

Out of these 20 sets, we selected 4 sets for En-Hi word ordering quality assessment. The evaluators were asked to rank these sets on the basis of word order quality without giving much emphasis on translation quality. The evaluators rated the sentences '0' if the word order was unacceptable (but still the sentence might be comprehensible in some cases) and '1' if the order was acceptable.

The reordering rules for our system were tested for both gold and automatic constituency parse based reordering accuracy. Table 3 shows the reordering results.

The reordering rules were also tested on gold

Input	(S (PP (IN In) (NP (CD 2006))) (, ,) (NP (DT the) (JJ Lebanese) (NNP Hezbollah) (NN militia)) (VP (VBD seemed) (ADJP (JJ able) (S (VP (TO to) (VP (VB rain) (NP (NNS rockets)) (PP (IN at) (NP (NN will)))) (PP (IN during) (NP (NP (DT a) (JJ monthlong) (NN conflict)) (PP (IN with) (NP (DT the) (JJ Jewish) (NN state)))))))))
Pada formation	(S 2006 In , the Lebanese Hezbollah militia (VP (VBD seemed) (ADJP (JJ able) (S (VP (VP (VB rain) to (NP (NNS rockets)) (PP will at)) (PP (NP (NP a monthlong conflict during) (PP (NP the Jewish state with)))))))))
RBMS1	(S 2006 In , the Lebanese Hezbollah militia (VP (VBD seemed) (ADJP (JJ able) (S (VP will at rockets rain to (PP (NP (NP a monthlong conflict during) (PP (NP the Jewish state with)))))))))
RBMS1	(S 2006 In , the Lebanese Hezbollah militia (VP (VBD seemed) (ADJP (JJ able) (S (NP (NP a monthlong conflict during) (PP the Jewish state with)) will at rockets rain to))))
RRMS2	(S 2006 In , the Lebanese Hezbollah militia (VP (VBD seemed) (ADJP (S (NP (NP a monthlong conflict during) (PP the Jewish state with)) will at rockets rain to able))))
RBMS1	(S 2006 In , the Lebanese Hezbollah militia (NP (NP a monthlong conflict during) (PP the Jewish state with)) will at rockets rain to able seemed)
RBMS4	2006 In , the Lebanese Hezbollah militia the Jewish state with a monthlong conflict during will at rockets rain to able seemed
lexical substitution	2006 meṃ, lebānī hijabullāha ladāke yahūdī rājyoṃ ke sātha mahīnā bhara lambe saṃgharṣa ke daurana icchānusāra rauketa barasāne meṃ sakṣama dikhe

Table 2: Illustration of reordering rules

Anusāraka with gold parse	67%
Anusāraka with automatic parse	65.5%
Google Translate	42%

Table 3: Reordering accuracy results evaluated by common people

and automatic constituency parse based input by the developers but the reordering scale for developers was between 0-2 as shown in Table 4. See the results in Figure 4.

0	unacceptable
1	non-fluent but acceptable
2	fluent/acceptable

Table 4: Reordering evaluation criteria for developers

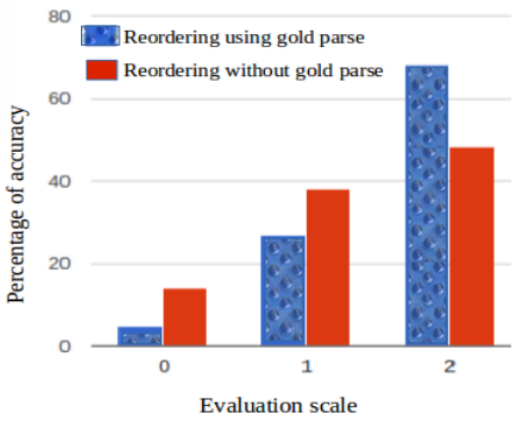

Figure 4: Percentage of reordering accuracy for En-Hi with and without gold parse

As expected, the results of the gold parse based word order are slightly better than the automatic parse based word ordering in case of developers' as well as common peoples' evaluation. The results confirm that a rule based reordering system performs better than a SMT based system.

#Sentences	#Words	#Phrases
100	2151	1637

Table 5: Test Corpus

We also report phrase reordering accuracy rather than sentence reordering accuracy, since our rules basically reorder phrases/padas and just because of incorrect reordering of one phrase, penalizing sentences of around 50 word length does not sound justified.

	#Phrases ordered correctly	#Phrases ordered incorrectly
Gold parse	97.5%	2.5%
Automatic parse	93%	7%

Table 6: Reordering results based on gold and automatic parse

Using our tool, Karan et al., (2014) have reported 21.84 BLEU score improvement over the baseline 20.04. After that the system has been improved. It should also be noted that their paper was not on reordering.

7 Error analysis

This section analyses various types of errors in TL reordering output.

Some cases rated as '0' were actually ordered correctly but the evaluators rated them low due to incorrect TL word substitutions. For instance, in 26, *reported* is translated as *sūcanā dī*, but in system generated output it was translated as *pāyā*.

(26) Hubbard reported from Cairo.
Hubbard.NOM sūcanā de.PT se Cairo
'Hubbard ne Cairo se sūcanā dī.'

In most of the cases, the adverbs like *now, mainly, likely*, etc. were misplaced as shown in 27. Or incorrect insertions of the words like *ki* (that), *isaliye* (hence), etc. were made in TL as shown in 28. For example,

(27) Saleh is **now** in the U.S. for further medical treatment.
'Saleh **aba** āge kī cikitsā ke liye U.S. meṃ hai'
'Saleh āge kī cikitsā ke liye **aba** U.S. meṃ hai'
'*Saleh āge kī cikitsā ke liye U.S. meṃ **aba** hai'

(28) Since chalk first touched slate, schoolchildren have wanted to know: What is on the test?

In 28, insertion of *taba se* (thenceforth) is required for a fluent and comprehensible sentence in Hindi.

We noticed an interesting case where the shared verb should be repeated while translation in Hindi. For instance, in 29, the shared verb *swim* makes two *tiṅanta* padas with each auxiliary/*tiṅ*, *can* and *help*. There also, if the shared verb is a conjunct verb, then the verbalizer/helping verb (Begum et al., 2011) behaves as a *dhātu* and makes an independent *pada* with the auxiliary as shown in bold in 30.

(29) She can and will **swim**.
vaha *BE ABLE TO* aura *FT* **taira**
'vaha **taira sakatī hai** aura **tairegī**'

(30) ... he **can** and **will**
... vaha *BE ABLE TO* aura *FT*
help fight the country's active
madada kara ladanā deśa kī sakriya
Al-Qaida branch.
Al-Qaida śākhā
'... vaha deśa meṃ sakriya Al-Qaida śākhā' se ladane meṃ **madada kara sakatā hai** aura **karegā**'

8 Conclusion and Future Work

This paper presented reordering rules for English-Hindi language pair using the concept of *pada* from Pāṇinian Grammar. We developed rules for generating fluent English-Hindi word order. We obtained better results than the SMT system Google Translate which show that the rules are accurate enough to enhance the translation quality. We elaborated on the concept of *pada* and presented a method for identification of *pada* and generation of reordering rules. We have also claimed that the approach presented in this paper is generic enough to be applied on English-Indian language MT systems.

Acknowledgments

We would like to express our special thanks to Prof. Vineet Chaitanya for his expert guidance and support. We would like to extend our thanks to Mrs. Sirisha Manju, Mrs. P Roja Laxmi, Mrs. Mahalaxmi, Mr. KRS Harsha, Ms. Pratibha Rani, Mr. Ganesh Katrapati and the evaluators for their help in implementing the rules, preparing the data, conducting the experiments and evaluation.

References

Amba Padmanathrao Anantpur. 2009. *Anusaaraka: An approach for MT taking insights from the Indian Grammatical Tradition*. Ph.D. thesis, University of Hyderabad.

Rafiya Begum, Karan Jindal, Ashish Jain, Samar Husain, and Dipti Misra Sharma. 2011. Identification of Conjunct verbs in Hindi and its effect on Parsing Accuracy. In *Computational Linguistics and Intelligent Text Processing*, pages 29–40. Springer.

Akshar Bharati and Amba Kulkarni. 2011. 'subject' in english is *abhihita*.

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

Akshar Bharati, Medhavi Bhatia, Vineet Chaitanya, and Rajeev Sangal. 1996. Paninian grammar framework applied to English. *Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur*.

Akshar Bharati, Rajni Moona, Smriti Singh, Rajeev Sangal, and Dipti Mishra Sharma. 2004. Mteval: an evaluation methodology for machine translation systems. In *Proc. SIMPLE Symp on Indian Morphology, Phonology and Lang Engineering*.

Akshar Bharati, Sukhada, Dipti M Sharma, and Soma Paul, 2015. *Sanskrit and Computational Linguistics*, chapter Anusāraka Dependency Schema from Pāṇinian Perspective. D. K. Publishers.

Niladri Chatterjee, Anish Johnson, and Madhav Krishna. 2007. Some improvements over the BLEU metric for measuring translation quality for Hindi. In *Computing: Theory and Applications, 2007. ICCTA'07. International Conference on Computing: Theory and Applications*, pages 485–490. IEEE.

Marta R Costa-Jussà and José AR Fonollosa. 2009. An Ngram-based reordering model. *Computer Speech & Language*, 23(3):362–375.

Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya. 2001. Interlingua-based English–Hindi machine translation and language divergence. *Machine Translation*, 16(4):251–304.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, pages 447–464.

Bonnie Jean Dorr. 1993. *Machine translation: a view from the Lexicon*. MIT press.

Kapildev Dvivedi. 1953. *Rachanānuvādakaumudī*. Vishwavidyalaya Prakashan.

Joseph C Giarratano and Gary Riley. 1998. *Expert systems*. PWS Publishing Co.

Yamuna Kachru. 2006. *Hindi*. London Oriental and African Language Library, John Benjamins.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

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.

Mahavir. 1984. *Samartha Theory of Pāṇini and Sentence Derivation*. Munshiram Manoharlal Publishers.

Raj Nath Patel, Rohit Gupta, Prakash B Pimpale, and Sasikumar M. 2013. Reordering rules for English-Hindi SMT.

Sandhya Rathore. 1998. Kaunda bhatta's vaiyakaranabhusanasara: an analytical study. *Indian Council of Philosophical Research (distributed by Munshiram Manoharlal, Delhi)*.

Smriti Singh, Mrugank Dalal, Vishal Vachani, Pushpak Bhattacharyya, and Om P Damani. 2007. Hindi generation from interlingua. *Machine Translation Summit (MTS 07), Copenhagen*.

Jag Deva Singh. 1991. *Pāṇini, his description of Sanskrit: An analytical study of Aṣṭādhyāyī*. Munshiram Manoharlal Publishers.

Karan Singla, Kunal Sachdeva, Diksha Yadav, Srinivas Bangalore, and Dipti Misra Sharma. 2014. Reducing the impact of data sparsity in statistical machine translation. *Syntax, Semantics and Structure in Statistical Translation*.

Prakya Sreesaila Subrahmanyam. 1999. *Pa:ṇinian Linguistics*. Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.