# Reordering rules for English-Hindi SMT

**Raj Nath Patel, Rohit Gupta, Prakash B. Pimpale and Sasikumar M**
CDAC Mumbai, Gulmohar Cross Road No. 9,
Juhu, Mumbai-400049
India
{rajnathp,rohitg,prakash,sasi}@cdac.in

## Abstract

Reordering is pre-processing stage for Statistical Machine Translation (SMT) system where the words of the source sentence are reordered as per the syntax of the target language. We are proposing a rich set of rules for better reordering. The idea is to facilitate the training process by better alignments and parallel phrase extraction for a phrase based SMT system. Reordering also helps the decoding process and hence improving the machine translation quality. We have observed significant improvements in the translation quality by using our approach over the baseline SMT. We have used BLEU, NIST, multi-reference word error rate, multi-reference position independent error rate for judging the improvements. We have exploited open source SMT toolkit MOSES to develop the system.

## 1 Introduction

This paper describes syntactic reordering rules to reorder English sentences as per the Hindi language structure. Generally in reordering approach, the source sentence is parsed(E) and syntactic reordering rules are applied to form reordered sentence(E`). The training of SMT system is performed using parallel corpus having source side reordered(E`) and target side. The decoding is done by supplying reordered source sentences. The source sentences prior to decoding are reordered using the same syntactic rules as applied for the training data. So, this process works as a preprocessing stage for the phrase-based SMT system. It has been observed that reordering as a pre-processing stage is beneficial for developing English-Hindi phrase based SMT system (Ramanathan et al., 2008; Rama et al., 2008). This paper describes a rich set of rules for the structural transformation of English sentence to Hindi language structure using Stanford (De et al., 2006) parse tree on source side. These rules are manually extracted based on analysis of source sentence tree and Hindi translation.

For the evaluation purpose we have trained and evaluated three different phrase based SMT systems using MOSES toolkit (Koehn et al. 2007) and GIZA++(Och and Ney, 2003). The first system was non-reordered baseline (Brown et al., 1990; Marcu and Wong, 2002; Koehn et al., 2003), second using limited reordering described in Ramanathan et al. (2008) and third using improved reordering technique proposed in the paper. Evaluation has been carried out for end to end English-Hindi translation outputs using BLEU score (Papineni et al., 2001), NIST score (Doddington, 2002), multi-reference position-independent word error rate (Tillmann et al., 1997), multi-reference word error rate (Nießen et al., 2000). We have observed improvement in each of these evaluation metrics used. Next section discusses related work. Section 3 describes our reordering approach followed by experiments and results in section 4 and conclusion in section 5.

## 2 Related Work

Various pre-processing approaches have been proposed for handling syntax within SMT systems. These proposed methods reconcile the word-order differences between the source and target language sentences by reordering the source prior to the SMT training and decoding stages. For English-Hindi statistical machine translation reordering approach is used by Ramanathan et al. (2008) and Rama et al. (2008). This approach (Ramanathan et al. 2008) has shown significant improvements over baseline (Brown et al., 1990; Marcu and Wong, 2002; Koehn et al., 2003). The BLEU score for the system has increased from 12.10 to 16.90 after reordering. The same reordering approach (Ramanathan et al., 2008) used by us has shown slight improvement in BLEU score of 0.64 over baseline i.e. BLEU score increased from 21.55 to



*Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 34–41,
Sofia, Bulgaria, August 8, 2013. 

22.19 compare to +4.8 BLEU point increase in the previous case. The reason can be, when the system is able to get bigger chunks from the phrase table itself the local reordering (within phrase) is not needed and the long distance reordering employed in the earlier approach will be helpful for overall better translation. It may not be able to show significant improvements when local reordering is not captured by the translation model.

Other language pairs have also shown significant improvement when reordering is employed. Xia and Mc-Cord (2004) have observed improvement for French-English and Chao et al. (2007) for Chinese-English language pairs. Nießen and Ney (2004) have proposed sentence restructuring whereas Collins et al. (2005) have proposed clause restructuring to improve German-English SMT. Popovic and Ney (2006) have also reported the use of simple local transformation rules for Spanish-English and Serbian-English translation.

Recently, Khalilov and Fonollosa (2011) proposed a reordering technique using deterministic approach for long distance reordering and non-deterministic approach for short distance reordering exploiting morphological information. Some reordering approaches are also presented exploiting the SMT itself (Gupta et al., 2012; Dlougach and Galinskaya, 2012).

Various evaluation techniques are available for reordering and overall machine translation evaluation. Particularly for reordering Birch and Osborne (2010) have proposed LRScore, a language independent metric for evaluating the lexical and word reordering quality. The translation evaluation metrics include BLEU (Papineni et. al., 2002), Meteor (Lavie and Denkowski, 2009), NIST (Doddington, 2002), etc.

## 3 Reordering approach

Our reordering approach is based on syntactic transformation of the English sentence parse tree according to the target language (Hindi) structure. It is similar to Ramanathan et al. (2008) but the transformation rules are not restricted to "SVO to SOV" and "pre-modifier to post-modifier" transformations only.

The idea was to come up with generic syntactic transformation rules to match the target language grammatical structure. The motivation came from the fact that if words are already in a correct place with respect to other words in the sentence, system doesn't need to do the extra work of reordering at the decoding time. This problem becomes even more complicated when system doesn't able to get bigger phrases for translating a sentence. Assuming an 18 words sentence, if system is able to get only 2 word length phrases, there are 362880(9!) translations (permutations) possible (still ignoring the case where one phrase having more than one translation options) for a sentence.

The source and the target sentences are manually analyzed to derive the tree transformation rules. From the generated set of rules we have selected rules which seemed to be more generic. There are cases where we have found more than one possible correct transformations for an English sentence as the target language (Hindi) is a free word order language within certain limits. In such cases word order close to English structure is preferred over possible word orders with respect to Hindi.

We identified 5 categories which are most prominent candidates for reordering. These include VPs (verb phrases), NPs (noun phrases), ADJPs (adjective phrase), PPs (preposition phrase) and ADVPs (adverb phrase). In the following subsections, we have described rules for these in more detail.

| Tag | Description(Penn tags) |
|---|---|
| *dcP* | A*ny, parser generated phrase* |
| *pp* | *Prepositional Phrase(PP)* |
| *whP* | *WH Phrase(WHNP, WHADVP, WHADJP, WHPP)* |
| *vp* | *Verb Phrase(VP)* |
| *sbar* | *Subordinate clause(SBAR)* |
| *np* | *Noun phrase(NP)* |
| *vpw* | *Verb words(VBN, VBP, VB,VBG, MD, VBZ, VBD)* |
| *prep* | *Preposition words(IN, TO,VBN,VBG)* |
| *adv* | *Adverbial words(RB, RBR, RBS)* |
| *adj* | *Adjunct word(JJ,JJR,JJS)* |
| *advP* | *Adverb phrase(ADVP)* |
| *punct* | *Punctuation(,)* |
| *adjP* | *Adjective phrase(ADJP)* |
| *OP* | *advP, np and/or pp* |
| *Tag** | *One or more occurrences of Tag* |
| *Tag?* | *Zero or one occurrence of Tag* |

Table 1: Tag description

The format for writing the rules is as follows:
*Type_of_phrase(tag1 tag2 tag 3: tag2 tag1 tag3)*

This means that *"tag1 tag2 tag3"*, structure has been transformed to *"tag2 tag1 tag3"* for the type_of_phrase. This type_of_phrase denotes our category (NP, VP, ADJP, ADVP, PP) in which rule fall. The table given above explains about various tags and corresponding Penn tags used in writing these rules.

The following subsections explain the reordering rules. The higher precedence rule is written prior to the lower precedence. In general the more specific rules have high precedence. Each rule is followed by an example with intermediate steps of parsing and transformation as per the Hindi sentence structure. "Partial Reordered" shows the effect of the particular rule whereas "Reordered" shows impact of the whole reordering approach. The Hindi (transliterated) sentence is also provided as a reference for the corresponding English sentence.

### 3.1 Noun Phrase Rules

*NP (np1 PP[ prep NP[ np2 sbar]] : np2 prep np1 sbar)* *(1)*

**English: The time of the year** when nature dawns all its colorful splendor, is beautiful.

**Parse:** *[NP* ***(****np1* the time***)*** *[PP (prep* of***)*** *[NP (np2* the year***)*** *(sbar* when nature dawns all its colorful splendor***)]]]*** , is beautiful .

**Partial Reordered:** ***(np2*** **the year*****)*** ***(prep*** **of*****)*** ***(np1*** **the time*****)*** *(sbar* when nature dawns all its colorful splendor*)* , is beautiful .

**Reordered:** *(np2* the year*) (prep* of*) (np1* the time*) (sbar* when nature all its colorful splendor dawns*)* , beautiful is .

**Hindi:** *varsh ka samay jab prakriti apne sabhi rang-birange vabahv failati hai, sundar hai* .

*NP(np SBAR[ S[ dcP ]] :dcP np)* *(2)*

**English:** September to march is **the best season to visit Udaipur**.

**Parse**: September to March is *[NP (np* the best season*) [SBAR [S (dcP* to visit Udaipur*)]]]* .

**Partial Reordered**: September to March is ***(dcP*** **to visit Udaipur*****)*** ***(np*** **the best season*****)*** .

**Reordered**: September to March *(dcP* Udaipur visit to*) (np* the best season*)* is .

**Hindi**: *september se march udaipur ghumane ka sabse achcha samay hai* .

*NP(np punct advP : advP punct np)* *(3)*

**English**: **The modern town of Mumbai, about 50 km south of Navi Mumbai** is Kharghar.

**Parse**: The modern town of *[NP (np* Mumbai*) (punct ,) (advP* about 50 km south of Navi Mumbai*)]* is Kharghar .

**Partial Reordered**: ***(advP*** **about 50 km south of Navi Mumbai*****)*** ***(punct ,)*** ***(dcP*** **The modern town of Mumbai*****)*** is kharghar .

**Reordered**: *(advP* Navi Mumbai of about 50 km south*) (punct ,) (dcP* Mumbai of the modern town*)* kharghar is .

**Hindi**: *navi mumbai ke 50 km dakshin me mumbai ka adhunic sahar kharghar hai* .

*NP( np vp : vp np)* *(4)*

**English**: The main attraction is a **divine tree called as 'Kalptaru'**.

**Parse**: The main attraction is *[NP (np* a divine tree*) (vp* called as 'Kalptaru'*) ]* .

**Partial Reordered**: The main attraction is ***(vp*** **` called as 'Kalptaru'*****)*** ***(np*** **a divine tree*****)*** .

**Reordered**: The main attraction *(vp* ` Kalptaru ' as called*) (np* a divine tree*)* is .

**Hindi**: *iska mukhya akarshan kalptaru namak ek divya vriksh hai* .

### 3.2 Verb Phrase Rules

*VP( vpw PP [ prep NP[ np punct? SBAR[ whP dcP ]]] : np prep vpw punct? whP dcP)* *(5)*

**English:** The best time to visit **is in the afternoon** when the crowd thins out.

**Parse:** The best time to visit *[VP (vpw* is*) PP[ (prep* in*) NP[ (np* the afternoon*) [SBAR (whP* when*) (dcP* the crowd thins out*)]]]* .

**Partial Reordered:** The best time to visit ***(np*** **the afternoon*****)*** ***(prep*** **in*****)*** ***(vpw*** **is*****)*** *(whP* when*) (dcP* the crowd thins out*)* .

**Reordered:** visit to The best time *(np* the afternoon*) (prep* in*) (vpw* is*) (whP* when*) (dcP* the crowd thins out*)* .

**Hindi:** *bhraman karane ka sabase achcha samay dopahar me hai jab bhid kam ho jati hai* .

*VP( vpw NP[ np punct? SBAR[ whP dcP ]] : np vpw punct? whP dcP)* *(6)*

**English**: Jaswant Thada **is a white marble monument** which was built in 1899 in the memory of Maharaja Jaswant Singh II.

**Parse**: jaswant thada *[VP (vpw* is*) [NP (np* a white marble monument*) [SBAR (whP* which*) (dcP* was built in 1899 in the memory of Maharaja Jaswant Singh II*)]]* .

**Partial Reordered:** Jaswant Thada ***(np*** **a white marble monument*****) (vpw*** **is*****)*** *(whP* which*) (dcP* was built in 1899 in the memory of Maharaja Jaswant Singh II*)* .

**Reordered:** Jaswant Thada *(np* a white marble monument*) (vpw* is*) (whP* which*) (dcP* Maharaja Jaswant Singh II of the memory in 1899 in built was*)* .

**Hindi**: *jaswant thada ek safed sangmarmar ka smarak hai jo ki maharaja jaswant singh dwitiya ki yad me 1889 me banwaya gaya tha* .

*VP(vpw OP sbar : OP vpw sbar ) (7)*

**English:** Temples in Bhubaneshwar are **built beautifully on a common plan** as prescribed by Hindu norms.

**Parse:** Temples in Bhubaneshwar are *[VP (vpw* built*) (advP* beautifully*) (pp* on a common plan*) (sbar* as prescribed by Hindu norms*)]* .

**Partial Reordered:** Bhubaneshwar in Temples are ***(advP* beautifully*) (pp* a common plan on*) (vpw* built*)*** *(sbar* as prescribed by Hindu norms) .

**Reordered:** Bhubaneshwar in Temples *(advP* beautifully*) (pp* a common plan on*) (vpw* built*)* are *(sbar* as Hindu norms by prescribed*)* .

**Hindi:** *bhubaneswar ke mandir hindu niyamon dwara nirdharit samanya yojana ke anusar banaye gaye hain* .

*VP(vpw pp1 pp*2: pp*2 pp1 vpw) (8)*

**English:** Avalanche is **located at a distance of 28 Kms from Ooty**.

**Parse:** Avalanche is *[VP (vpw* located*) (pp1* at a distance of 28 kms*) (pp2* from Ooty*)]* .

**Partial Reordered:** Avalanche is ***(pp2* from Ooty*) (pp1* at a distance of 28 kms*) (vpw* located*)*** .

**Reordered:** Avalanche *(pp2* Ooty from *) (pp1* 28 kms of a distance at*) (vpw* located*)* is .

**Hindi:** *avalanche ooty se 28 km ki duri par sthit hai* .

*VP(vpw np pp : np pp vbw) (9)*

**English:** Taxis and city buses available outside the station, **facilitate access to the city**.

**Parse:** Taxis and city buses available outside the station , *[VP (vpw* facilitate*) (np* access*) (pp* to the city*)]* .

**Partial Reordered:** Taxis and city buses available outside the station **, *(pp* to the city*) (np* access*) (vpw* facilitate*)*** .

**Reordered:** Taxis and city buses the station outside available , *(pp* the city to*) (np* access*) (vpw* facilitate*)* .

**Hindi:** *station ke baahar sahar jane ke liye taksi aur bus ki suvidha upalabdha hai* .

*VP ( prep dcP : dcP prep) (10)*

**English:** A wall was built **to protect it**.

**Parse:** A wall was built *[VP (prep* to*) (dcP* protect it*)]* .

**Partial Reordered:** A wall was built ***(*protect it*) (prep* to*)*** .

**Reordered:** A wall *(dcP* it protect*) (prep* to*)* built was .

**Hindi:** *ek diwar ise surakshit karane ke liye banayi gayi thi* .

*VP(adv vpw dcphrase: dcphrase adv vpw) (11)*

**English:** Modern artist such as French sculptor Bartholdi is **best known by his famous work**.

**Parse:** Modern artists such as French sculptor Bartholdi is *[VP (adv* best*) (vpw* known*) (dcP* by his famous work*)]* .

**Partial Reordered:** Modern artists such as French sculptor Bartholdi is ***(dcP* by his famous work*) (adv* best*) (vpw* known*)*** .

**Reordered:** such as French sculptor Bartholdi Modern artists *(dcP* his famous work by*) (adv* best*) (vpw* known*)* is .

**Hindi:** *french shilpkar bartholdi jaise aadhunik kalakar apane prashidha kam ke liye vishesh rup se jane jate hain* .

*VP(advP vpw dcP: advP dcP vpw) (12)*

**English:** Bikaner, popularly **known as the camel county** is located in Rajasthan.

**Parse:** Bikaner , *[VP (advP* popularly*) (vpw* known*) (dcP* as the camel country*)]* is located in Rajsthan .

**Partial Reordered:** Bikaner **,** *(advP* popularly*) **(dcP* as the camel country*) (vpw* known*)*** is located in Rajsthan .

**Reordered:** Bikaner , *(advP* popularly*) (dcP* the camel country as*) (vpw* known*)* Rajsthan in located is .

**Hindi:** *bikaner , jo aam taur par unton ke desh ke naam se jana jata hai, rajasthan me sthit hai* .

*VP(vpw adv? adjP? dcP: dcP adjP? adv? vpw) (13)*

**English:** This palace has **been beautiful from many years**.

**Parse:** This palace has *[VP (vpw* been*) (adjP* beautiful*) (dcP* from many years*)]* .

**Partial Reordered:** This palace has ***(dcP* from many years*) (adjP* beautiful*) (vpw* been*)*** .

**Reordered:** This palace *(dcP* many years from*) (adjP* beautiful*) (vpw* been*)* has .

**Hindi:** *yah mahal kai varson se sunder raha hai* .

### 3.3 Adjective and Adverb Phrase Rules

*ADJP( vpw pp : pp vpw )* *(14)*

**English:** The temple is **decorated with paintings depicting incidents**.

**Parse:** The temple is *[ADJP (vpw* decorated*) (pp* with paintings depicting incidents *)]* .

**Partial Reordered:** The temple is ***(pp* with paintings depicting incidents*) (vpw* decorated*)*** .

**Reordered:** The temple *(pp* incidents depicting paintings with*) (vpw* decorated*)* is .

**Hindi:** *mandir ghatnao ko darshate hue chitron se sajaya gya hai* .

*ADJP( adjP pp : pp adjP )* *(15)*

**English:** As a resul, temperatures are now higher than ever before .

**Parse**: As a result , temperatures are now *[ADJP (adjP* higher*) (pp* than ever*)]* before .

**Partial Reordered:** As a result , temperatures are now *(pp* than ever*) (adj* higher*)* before .

**Reordered:** a result As , temperatures now before ***(pp* ever than*) (adj* higher*)*** are .

**Hindi:** *parinam swarup taapman ab pahle se bhi adhik hai* .

*ADJP( adj dcP : dcP adj )* *(16)*

**English:** The Kanha National park is **open to visitors**.

**Parse:** The Kanha National park is *[ADJP (adj* open*) (dcP* to visitors*)]* .

**Partial Reordered:** The Kanha National park *is* ***(pp to* visitors *) (adj* open*)*** .

**Reordered:** The Kanha National park *(pp* visitors to*) (adj* open*)* is .

**Hindi:** *kanha national park paryatakon ke liye khula hai* .

*ADVP( adv dcP: dcP adv )* *(17)*

**English:** The temple is most favored spot for tourists **apart from the pilgrims**.

**Parse:** The temple is most favored spot for tourists *[ADVP (adv* apart*) (dcP* from the pilgrims*)]* .

**Partial Reordered:** The temple is most favored spot for tourists ***(dcP* from the pilgrims *) (adv* apart*)*** .

**Reordered:** The temple most favored spot *(dcP* the pilgrims from*) (adv* apart*)* is .

**Hindi:** *mandir teerth yatriyon ke alawa paryatkon ke liye bhi lokpriya sthal hai* .

### 3.4 Preposition Phrase Rules

*PP( adv prep? dcP : dcP prep? adv )* *(18)*

**English:** Does kalajar occur **because of sun**?

**Parse:** Does kalajar occur *[PP (adv* because*) (prep?* of*) (dcP* sun*)]* ?

**Partial Reordered:** Does kalajar occur ***(dcp* sun*) (prep?* of*) (adv* because*)*** ?

**Reordered:** Does kalajar *(dcp* sun*) (prep?* of*) (adv* because*)* occur*?*

**Hindi:** *kya kalajar dhup ke karan hota hai ?*

| | |
|---|---|
| input | Ahmedabad was named after the sultan Ahmed Shah, who built the city in 1411. |
| baseline | ahmedabad was named after the sultan ahmed shah, who built the city in 1411.<br>अहमदाबाद के नाम पर रखा गया सुल्तान अहमद shah, वाले शहर 1411.<br>*ahamdabad ke nam par rakha gaya sultan ahamad shah, wale shahar 1411.* |
| limited reordering | ahmedabad the sultan ahmed shah , who the city 1411 in built after named was .<br>अहमदाबाद का नाम सुल्तान अहमदशाह के , जिसने १४११ में शहर बनवाया के नाम पर रखा गया था ।<br>*ahamdabad ka nam sultan ahamadshah ke , jisane 1411 me shahar banawaya ke nam par rakha gaya tha .* |
| our approach | ahmedabad the sultan ahmed shah after named was , who 1411 in the city built .<br>अहमदाबाद का नाम सुल्तान अहमदशाह के नाम से पड़ा था जिसने १४११ में शहर बनवाया था ।<br>*ahamadabad ka nam sultan ahamadshah ke nam se pada tha jisane 1411 me shahar banawaya tha .* |
| reference | अहमदाबाद का नाम सुल्तान अहमदशाह के नाम पर पड़ा था, जिसने १४११ में शहर बनवाया था ।<br>*ahamadabad ka nam sultan ahamadshah ke nam par pada tha jisane 1411 me shahar banawaya tha .* |

Table 2: Comparison of translation on a sentence from test corpus

## 4 Experiments and Results

The experiments were carried out on the corpus described in Table 3 below.

| | #Sentences | #Words |
|---|---|---|
| Training | 94926 | 1235163 |
| Tuning | 1446 | 23600 |
| Test | 500 | 9792 |

Table 3: Corpus distribution

The baseline system was setup by using the phrase-based model (Brown et al., 1990; Marcu and Wong, 2002; Koehn et al., 2003). For the language model, we carried out experiments and found on comparison that 5-gram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998) to be the best performing. Target Hindi corpus from the training set was used for creating the language model. The KenLM (Heafield., 2011) toolkit was used for the language modeling experiments. The tuning corpus was used to set weights for the language models, distortion model, phrase translation model etc. using minimum error rate training (Och, 2003). Decoding was performed using the MOSES decoder. Stanford constituency parser (De et al., 2006) was used for parsing.

Table 2 above describes with the help of an example how the reordering and hence the translation quality has improved. From the example it can be seen that the translation by system using our approach is better than the other two systems. The output translation is structurally more correct in our approach and conveys the same meaning with respect to the reference translation.

| **phrase-length** | **#phrases** | | | **#distinct-phrases(distinct on source)** | | |
|---|---|---|---|---|---|---|
| | **baseline** | **limited reordering/ %IOBL/ IOBL** | **our approach/ %IOBL/ IOBL** | **baseline** | **limited reordering/ %IOBL/ IOBL** | **our approach/ %IOBL/ IOBL** |
| 2 | 537017 | 579878/ 7.98/ 42861 | 579630/ 9.98/ **42613** | 208988 | 249847/ 19.55/ 40859 | 254393/ ***21.72***/ **45405** |
| 3 | 504810 | 590265/ 16.92/ 85455 | 616381/ 22.10/ **111571** | 292183 | 384518/ 31.62/ 92335 | 408240/ ***39.72***/ **116057** |
| 4 | 406069 | 493637/ 21.56/ 87568 | 531904/ 30.98/ **125835** | 268431 | 372282/ 38.68/ 103851 | 409966/ ***52.72***/ **141535** |
| 5 | 313368 | 391490/ 24.92/ 78122 | 431135/ 37.58/ **117766** | 221228 | 313723/ 41.80/ 92495 | 354273/ ***60.13***/ **133045** |
| 6 | 231146 | 292899/ 26.71/ 61753 | 327192/ 41.55/ **96046** | 170852 | 244643/ 43.19/ 73791 | 279723/ ***63.72***/ **108871** |
| 7 | 154800 | 196679/ 27.05/ 41879 | 220868/ 42.67/ **66068** | 119628 | 170108/ 42.19/ 50480 | 194881/ ***62.90***/ **75253** |

Table 4: Phrase count analysis

The Table 5 below lists four different evaluations of the systems under study. For BLEU and NIST higher score is considered as better and for mWER and mPER lower score is desirable. Table 5 shows the results of comparative evaluation of baseline, limited reordering and our approach with improved reordering. We find that addition of more reordering rules show substantial improvements over the baseline phrase based system and the limited reordering system (Ramanathan et al., 2008). The impact of improved syntactic reordering can be seen as the BLEU and NIST scores have increased whereas mWER and mPER scores have decreased.

| | **BLEU** | **NIST** | **mWER %** | **mPER %** |
|---|---|---|---|---|
| baseline | 21.55 | 5.72 | 68.08 | 45.54 |
| limited reordering | 22.19 | 5.74 | 66.44 | 44.70 |
| our approach | **24.47** | **5.88** | **64.71** | **43.89** |

Table 5: Evaluation scores

Table 4 above shows the count of overall phrases and distinct phrases (distinct on source) for baseline, limited reordering approach and our improved reordering approach. The table also shows increase over baseline (IOBL) and percentage increase over baseline(%IOBL) for limited reordering and improved reordering. We have observed that no. of distinct phrases extracted from the training corpus get increased. The %IOBL for bigger phrases is more compare to shorter phrases. This can be attributed to the better alignments resulting in extraction of more phrases (Koehn et al., 2003).

We have also observed that the overall increase is even lesser than the increase in no. of distinct phrases (distinct on source) for all the phrase-lengths in our approach (e.g. 42613 and 45405 for phrase-length 2) which shows that reordering makes word alignments more consistent and reduces multiple entries for the same source phrase. The training was done on maximum phrase-length 7(default).

## 5 Conclusion

It can be seen that addition of more reordering rules improve translation quality. As of now we have tried these rules only for English-Hindi pair, but the plan is to employ similar reordering rules in other English-Indian language pairs as most Indian languages are structurally similar to Hindi. Also plans are there to go for comparative study of improved reordering system and hierarchical model.

## References


Alexandra Birch , Miles Osborne and Phil Blunsom. 2010. Metrics for MT evaluation: evaluating reordering. *Machine Translation* 24, no. 1: 15-26.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational linguistics* 16(2): 79–85.

Wang Chao, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).*

Stanley F. Chen, Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.*

Marneffe De, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, vol. 6, pp. 449-454.

Jacob Dlougach and Irina Galinskaya. 2012. Building a reordering system using tree-to-string hierarchical model. In *Proceedings of the First Workshop on Reordering for Statistical Machine Translation at COLING*, Mumbai, India.

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*. Morgan Kaufmann Publishers Inc.

Rohit Gupta, Raj N. Patel and Ritesh Shah. 2012. Learning Improved Reordering Models for Urdu, Farsi and Italian using SMT. In *Proceedings of the first workshop on Reordering for Statistical Machine Translation, COLING 2012,* Mumbai, India.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics.*

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan 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.

Daniel Marcu, and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. Proceedings of EMNLP.

Sonja Nießen, Franz J. Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. International Conference on Language Resources and Evaluation.

Franz J. Och, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, Volume 29, number 1:19-51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*-Volume 1:pp. 160-167.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report, Thomas J. Watson Research Center.*

Taraka Rama, Karthik Gali and Avinesh PVS. 2008. Does Syntactic Knowledge help English-Hindi SMT ?. *Proceedings of the NLP Tools contest, ICON.*

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah, and M. Sasikumar. 2008. Simple syntactic and morphological processing can help English-Hindi statistical machine translation. *In International Joint Conference on NLP (IJCNLP08).*

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alex Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. *In European Conf. on Speech Communication and Technology.*

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 508. Association for Computational Linguistics.