Distance-based Reordering in English to Hindi Statistical Machine Translation

Sudhakar Kumawat
Department of Computer Engineering
Indian Institute of Technology, (BHU) Varanasi, India

Nitish Chandra
Department of Computer Engineering
Indian Institute of Technology,(BHU)Varanasi, India

ABSTRACT
This paper compares different reordering models on English to Hindi statistical machine translation system. The two Indo-European languages differ significantly in their word order preferences. While English follows SVO model, Hindi follows SOV model. Therefore both long distance and short distance reordering becomes important. The reordering models available in MOSES SMT are discussed and compared with a more novel approach called distance-based reordering. This new approach significantly improves the quality of English to Hindi translation, both in terms of BLEU score and subjective human evaluation..

General Terms
Machine Translation.

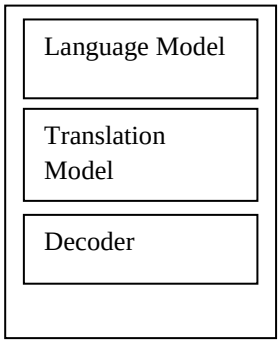
Keywords
Distance-based reordering, Statistical machine translation

1. INTRODUCTION
The two Indo-European languages, English and Hindi although belonging to the same family, differ significantly in their word order, syntax and morphology [1]. These differences between languages make the task of translation between source and target languages very difficult. Thus, linguistic differences along with the size of parallel corpora and length of sentences play an important role in machine translation [1][2].

All languages of the world are highly inflected which means their word order and forms change when their way of using in the sentence is changed. However, English have a simple inflection system as compared to Hindi where nouns, verbs and adjectives are inflected according to number, gender, case etc. In fact, Hindi is very rich in morphology [2]. This makes the task of a plain phrase-based statistical translation system very difficult as it may not be able to cope the differences in grammars of the languages correctly [1][2]. This paper studies and experiment the ways of improving the translation quality of a pair of language by making the two languages structurally similar at the pre-processing stage.

2. OVERVIEW OF THE SMT SYSTEM
Statistical machine translation (SMT) system is one of the applications of Noisy Channel Model. The noisy channel model of a SMT system for translating from Language 'S' to Language ' T' works as follows : The channel receives the input sentence 's' of language S, transforms it ("add noise") into the sentence 't' of Language T and sends 't' to a decoder. The decoder then determines the sentence 's' of language S that t is most likely to have arisen from and which is not necessarily identical to 's' [2][3].

Thus, for translating from language 'S' to language 'T' the SMT system requires three major componenets. A component called Language Model for computing probabilities to generate sentence 't', another component called Translation Model for computing translation probabilities of sentence 's' given 't', and finally, a component called Decoder for searching among possible sentences 's' for the one that gives the maximum value for P(s|t)P(t) [2][3].



2.1 Language Model
The goal of the SMT system is to estimate the probability of translation from source language to target language. In order to do this, a sentence is broken down into the product of conditional probabilities. The language model computes the probability of a word given it's preceding words in the sentence. This model is known as n-gram model [4][12]. For example, the probability of sentence 'S' is decomposed into probabilities of individual words 'w' as follows:

$P(s) = P(w_1, w2, w3 \ldots wn)$

$= P(w_1)\ P(w_2 \mid w_1)\ P(w_3 \mid w_1\ w_2)\ P(w_4 \mid w_1\ w_2\ w_3) \ldots..\ P(w_{n-1} \mid w_1\ w2 \ldots w_n)$

In order to calculate sentence probability, it is required to calculate the probability of a word, given the sequence of word preceding it. An n-gram model simplifies the task by approximating the probaility of a word given all the previous words.

Consider the following training set of data:

1. Ganga is a river.
2. It is a long river.
3. Water of Ganga is very pure.

Probabilities for bigram model are as shown below:

1. P (Ganga/is) = 0.6 P (is/a) = 1.0 P (a/river) = 0.4
2. P (it/is) = 1.0 P (a/long) = 0.1 P (long/river) = 0.2
3. P (water/of) = 0.2 P (of/Ganga) = 0.5 P (is/very) = 0.2 P(very/pure)= 0.4

The probability of a sentence: "Ganga is a long and pure river", can be computed as follows:

P (Ganga/is) * P (is/a) * P (long/a) * P (and /pure) *P (pure/river)

= 0.6 * 1.0 * 0.1 * 0.3 * 0.4

= 0.0072

2.2 Translational Model

This component of the SMT system calculates the conditional probability P(T/S) and is trained from the source-target language parallel corpus. The probabilities are calculated at the words and phrases level and not at the sentence level. This is why; the translated text is thought of as being produced from source word by word [4][5].

For example, denoting 'S' as source language text and 'T' as it's translated text in target language, together represented as (T/S).Using this notation, the translation is represented as:

(Ram talaab mein kud gaya | Ram jumped into the pond)

राम तालाब में कूद गया | ram jumped into the pond

One of the possible alignment for this pair of sentences are represented as :

राम तालाब में कूद गया | ram(1) jumped(4) into(3) the(3) pond(2)

For a source language text of length 'm' and it's translated text in target language of length 'l', the total number of possible alignments are 'lm'.

Here, in the above example word by word alignments are considered and denoted as A(S,T).All alignments have equally likely , therefore word order in 'T' and 'S' do not effect P(T/S) and likelihood of (T/S) , in terms of conditional probability P(T,a/S) is denoted as:

P(S/T) = sum P (S.a/T) [4]

The sum is over the elements of alignment set, A(S, T).English word has only exactly one connection. For the alignment,

P (राम तालाब में कूद गया | ram jumped into the pond), can be computed by multiplying the translation probabilities T(राम | ram(1)), T (तालाब | pond(2)), T(में|into(3)), T(null | the(3)) and T (कूद | jumped(2)).

To generate target sentence from source sentence, we have to follow the steps as given below [4][5]:

I. Select the length of S with probability L where L = P [length(S) = m] is a constant, i.e., all lengths are assumed to be equally likely with probability L.

II. Select an alignment with probability P (a/S). There are (l+1) possible alignments. Assuming all possible alignments are equally likely, the probability of alignment a,P(a/S), is as shown : $P(a/S)=L * 1/(l+1)^m$

III. Select the jth ^{English} word with a probability.

The joint likelihood of Hindi string and an alignment given an English string is given as:

P(S, a/T) = P (a/T) * P(S/a, T) [4]

T is the probability of seeing S_J in source sentence, given T_{aj} in target sentence. The alignment is determined by specifying the values of a_j for j from 1 to m, each of which can take value from 0 to l.

2.3 Decoder

The decoder component of SMT maximizes the probability of translated text. The words chosen by the decoder have the maximum probability of being in the translated text. Sentence 'T' is searched to maximize P(S/T) [4][5] i.e.:

P(S, T) = max [P (T) P(S/T)]

However, the problem here is searching the infinite space. Therefore, stacked search is used which maintain a list of partial alignment hypothesis. Here search starts with a null hypothesis and the translated text is obtained from a text of source language words that are not known [4][5]. For example:

(राम तालाब में कूद गया|∗), here * represents the null hypothesis, i.e. a unknown sequence of source language words. As the search proceeds, additional words are added to this hypothesis. The example below shows the process of adding words during search process:

(राम तालाब में कूद गया | ram (1)),

(राम तालाब में कूद गया | ram (1) pond(2)),

(राम तालाब में कूद गया | ram (1) pond (2) into (3) the(3) jumped (4)),

(राम तालाब में कूद गया | ram(1) into(3) the (3) pond (2) jumped (4)),

(राम तालाब में कूद गया | ram (1) jumped (4) into(3) the(3) pond(2))

The search process terminates when a complete alignment is found which is better than any of the incomplete alignments [4].

3. DATA AND THEIR PREPROCESSING

This section provides a brief overview of the data used in this study. We also summarize some statistics over our corpora.

We normalized all Hindi texts to make them usable for training of the translation system. We collected four different corpra of atleast three different domains from various sources. In addition, collected a large monolingual corpus from the web[3].

3.1 Parallel Data

The following four English-Hindi parallel corpora were used :

1. EMILLE is a 63 million word corpus of Indic languages which is distributed by European Language Resources Association (ELRA). The detail of Emille corpus available from their online manual.

2. English-Hindi parallel corpora available on LetsMT! Website.

3. English- Hindi parallel corpora on history of Delhi available online

4. Bhagvad Gita corpus available online.

Table 1: English parallel corpus size information.

Corpus	Source	SentPairs	Tokens	Vocabulary	sentence	Length
Emille	ELRA	7,957	210,597	5,969	26.47	9.77
Lets MT	web	8,736	153,519	9,087	17.57	9.87
Delhi	web	6,414	22,603	8,135	39.38	28.59
Gita	web	6,215	161,294	13,826	25.95	12.46

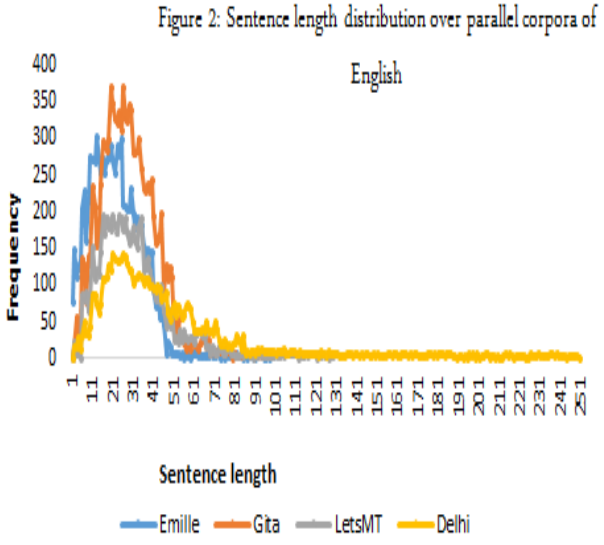
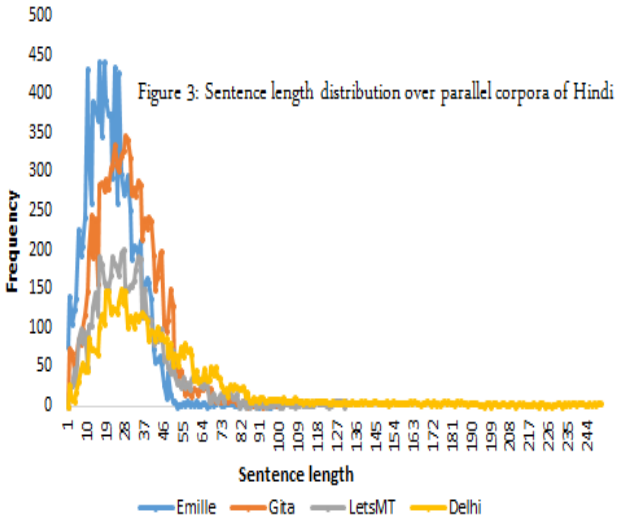


Table 2: Hindi parallel corpus size information.

Corpus	Source	SentPairs	Tokens	Vocabulary	Sentence	Length
Emille	ELRA	7,957	203,927	6,980	25.62	9.36
Lets MT	web	6,414	269,991	7,183	42.09	30.33
Delhi	web	6,215	185,690	12,457	29.88	14.44
Gita	web	8,736	200,179	9,626	22.91	13.07



4. REORDERING MODELS

This section addresses the problems that are specific to the English-Hindi language pair. Improvement techniques are also proposed to help the SMT system to solve these problems.

However, the focus of this paper is limited to the word order differences between English and Hindi language pair. As discussed earlier, English is SVO (subject->verb->object) language while Hindi is SOV language[3][6]. Therefore for high quality translations, the SMT system may have to perform short distance along with long distance reordering. Unfortunately long distance reordering have very high time and space complexity as there are too many partial hypothesis possible. The SMT system may have to terminate its search prematurely thus, losing a good partial hypothesis at initial stage[3].

Thus, this paper proposes a better approach by trying to make the word order of English text close to the expected word order of Hindi text at the pre-processing stage.

4.1 Lexical Reordering in Moses

MOSES system learns different reordering probabilities for each phrase during the training process. These probabilities are then conditioned on the lexical value of the phrase in the sentence. Therefore, such reordering models are also referred to as lexical reordering model [7]. In unidirectional reordering model, MOSES learns reordering probability of a phrase with respect to the previous phrase. Three reordering types (M, S, D) are included in MSD-unidirectional model [3]:

Monotone (M) - In this reordering type, the reordering of the target phrases is identical to the reordering of their counterparts in the source language [3][5][6].

Swap(S) - In this reordering type, the ordering of the two phrases is swapped in the target language, i.e. the preceding target phrase translates the following source phrase [3][5][6].

Discontinuous (D) - It means anything elsewise the source counterpart of the preceding target may lie before or after the counterpart of the current phrase but in neither case is the two source phrase adjacent [3][5][6].

4.2 Distance Based Reordering in Moses

Reordering of the target output phrases is modeled through relative distortion probability distribution d ($start_i$, end_{i-1}), where start i refers to the starting position of the source phrase that is translated into (i-1)th target phrase. The reordering distance is coumputed as ($start_i$ – end_{i-1}) [3][9].

The reordering distance is the number of words skipped (either forward or backward) when taking source words out of sequence. If two phrases are translated in sequence , then $start_i = end_{i-1} + 1$; i.e, the position of the first word of phrase i immediately follows the position of the last word of previous phrase. In this case, a reordering cost of d(0) is applied. Distance-based model gives linear cost to reordering distance i.e. movements of phrases over large distances are more expensive [3][9].

5. EXPERIMENTS AND RESULTS

The MOSES [11] baseline model setup is a plain phrase-based translation model combined with bidirectional reordering model while the distance-based pre-processing technique use both the bidirectional and distance based reordering models.
All experiments have been performed on normalized target data. All Hindi data have been normalized, i.e. training data, testing data and reference translations of development.
The compared BLEU [11] scores of MOSES baseline and distance-based systems are shown in Table 3. For all the corpora, the distance based system gave significantly better results than the baseline system.

Table 3: BLEU Scores (computed against one reference translation)

Parallel Data	Baseline	Distance -based
Emille	23.01	25.15
LetsMT	19.80	23.75
Delhi	13.90	16.76
Gita	13.56	14.67

6. HUMAN EVALUATOIN

The BLEU scores of the distance-based system were higher than the MOSES baseline system,
However, for language pair like English-Hindi, an computerized evaluation metric like BLEU may not give optimum results. Therefore subjective human evaluation of translation quality produced by the system becomes important. However, due to time and labour constraints the subjective evaluation could be done only on a limited set of data [3][12].

A limited set of test data of about 500 sentences was taken from which 120 sentences were selected randomly and fed to the system for translation and the translated Hindi text was presented to a native speaker of Hindi who was asked to assign to each Hindi translation one of the following three scores

0: Useless translation, even broad English meaning cannot be estimated.

1: Partial English meaning can be interpreted.

2: Correct and understandable translation may not be completely correct.

Here also the distance-based system gave better results than the baseline system.

Table 4: Human Evaluation

Category	Reference	Baseline	Distance-based
0	1	20	21
1	4	20	24
2	45	10	11

6. CONCLUSION

This paper compares different reordering models on English to Hindi statistical machine translation system. At first, significant amount of parallel and monolingual data from different domains was collected and normalized. Thereafter keeping focus on word order differences the Moses baseline system and distance-based pre-processing technique were compared. Experimental results show that the distance-based system outperformed the baseline model on all corpuses both in terms of BLEU score, the automatic evaluation system as well as subjective human judgements.

7. REFERENCES

[1] Bharati, Akshar, Vineet Chaitanya, and Rajeev Sangal. Natural Language Processing, a Paninian Perspective. Prentice Hall of India, 1995.

[2] Bojar, Ond, Pavel Stra, and Daniel Zeman. English-Hindi translation in 21 days. In Proceedings of the 6th International Conference on Natural Language Processing (ICON-2008) NLP Tools Contest, 2008.

[3] Bushra Jawaid, Daniel Zeman. Word-Order Issues in English-to-Urdu . PBML april 2011. http://ufal.mff.cuni.cz/~jawaid/publications/art-jawaid-zeman.pdf

[4] Nakul Sharma, P Bhatia, V Singh. English to Hindi Statistical Machine Translation System. Thapar University. 2011

[5] Koehn, Philipp. Statistical Machine Translation. Cambridge University Press, Cambridge, UK, 2010.

[6] Michel Galley, Christopher D. Manning. A Simple and Effective Hierarchical Phrase Reordering Model. Proceedings of the 2008 Conference Empirical Methods in Natural Language Processing .Honolulu, October 2008.

[7] Wang Ling, Joao Grac¸a, David Martins de Matos, Isabel Trancoso, Alan Black. Discriminative Phrase-based Lexicalized Reordering Models using weighted Reordering Graphs. Carnegie Mellon University, Pittsburgh, PA, USA.

[8] Jurafsky, Daniel and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall, Upper Saddle River, NJ, 2000. ISBN 0-13-095069-6.

[9] Kneser, Reinhard and Hermann Ney. Improved backing-off for m-gram language modeling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Los Alamitos, California, USA, 1995. IEEE Computer Society Press.

[10] Yizhao Ni, Distance phrase reordering for MOSES. Pattern Analysis and Intelligent Systems Research Group.Department of Engineering Mathematics University of Bristol

[11] Chen, Stanley F. and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In Technical report TR-10-98, Computer Science Group, Harvard, MA, USA, August 1998. Harvard University. URL http://research.microsoft.com/en-us/um/people/joshuago/tr-10-98.pdf.

[12] MOSES , GIZA ++,BLEU tool http://statmt.org/.