A Study on Reordering Methodologies for English-Punjabi Machine Translation

Shishpal Jindal
Ph.D. Research Scholar, I.K.G.
Punjab Technical University,
Jalanadhar, Punjab
shishpaldav@gmail.com

Vishal Goyal
Department of Computer
Science, Punjabi University,
Patiala, Punjab
vishal.pup@gmail.com

Jaskarn Singh Bhullar
Department of Applied Sciences,
MIMIT, Malout, Punjab
bhullarjaskarn@gmail.com

Abstract—**One of the challenging issues in Machine Translation is how to place the translated words in such order that they fit according to the target language. Both, English and Punjabi have syntactically different structure i.e. Subject-Verb-Object (SVO) and Subject-Object-Verb (SOV), respectively. Both English and Punjabi have relatively restrictive word orders. On the major challenge in the Machine Translation is the different word order between the source sentences to a target sentence. So, the process of reordering is required for decoding to skip words and cover them after translating later in the sentence. In this paper, we develop a methodology to handle the fundamental problems of word ordering for Statistical Machine Translation (SMT). We focus on evaluating SMT in general and the fact of reordering in particular. In other words we aim to investigate the viability of SMT in performing bilingual translation. Typological disparity of languages leads to wrong translation by monotone SMT systems. So, there is a need of reordering models with the capability to overcome the problem of this parity.**

Keywords- machine translation; SMT; reordering; SVO; SOV.

I. Introduction

The reordering problem in SMT originates from the fact that not all the words in a sentence can be back-to-back translated. This means some words skipped and be translated out of order in the source sentence to produce grammatically correct sentence in the target language. The main reason is that, reordering is needed. We propose to lessen the word order challenge including morpho-syntactical and statistical information in the context of a pre-translation reordering framework. A quite popular class of pre-ordering algorithm is a validation of the source part of parallel corpus before to translation. The first work on this approach is described in NeiBen and Ney [1], where morpho-syntactic information was used to account for the purpose of reordering. The main focus of this research paper is to handle the problem of reordering in SMT because the core role of reordering in Machine Translation affects several aspects of translation. It has been noticed that better reordering directly influence the quality of outcome. Additionally, the reordering method has a significant impact on determining the performance of translation system in terms of speed. The main aim of this paper is to present various handcrafted reordering rules for English to Punjabi machine translation. A monotone SMT system can effectively cope up different word order between the languages pair (English and Punjabi). There are few issues in English-Punjabi languages reordering as briefly discussed as follows.

- English is highly positional language with primitive morphology and default sentence structure as SVO.
- Punjabi is highly inflectional, with a rich morphology and default sentence structure as SOV.
- English used pre-positions while Punjabi uses post-positions.
- Punjabi language allows greater word order freedom.
- Punjabi language is a richer case marking system.

This paper is organized into five sections. Introduction has been presented in Section 1. Section 2, presents the related work. Proposed methodologies are discussed in Section 3. Reordering rules are discussed in Section 4. Finally, in Section 5, we have concluded the work.

II. Related Work

There are two ways to handle the problem of reordering. First way directly improves the reordering model insight the SMT system and the other by pre-reordering of the source sentences that is according to the target sentence. Many SMT models implement the brute force approach, introducing several reordering constraints. These constraints are not lexicalized. The world class reordering patterns were part of Och’s alignment template system [2]. The modern state-of-the-art phrase-based translation system Moses along with a distance based distortion model [3], implements the phrase based reordering [4]. Other researchers tried to include reordering rules which either have been defined manually [5] or have been learned statistically from the reordering patterns, in the parallel training data [6]. The distance-base penality model is used in MOSES [7]. In additionally a maximum entropy reordering model is used to predict the orientation of a phrase [8]. Previous work on English and Punjabi machine translation is limited. The rule based machine translation system discussed is one of the existing systems that translate from English text to Punjabi text. The objective of present work is to reorder the source sentence in such a way that source and target chunks become monotone. Other approaches were introduced that used more linguistic knowledge, for example the use of bi-text grammars that allow parsing the source and target languages. In our approach, we follow the idea of using a parallel training corpus with tagged source side to extract rules which allow a reordering before translation task. Moreover we use lexical information for some part of speech rules to solve ambiguity problem. By doing this we hope to handle the reordering problem by the proposed rules in this paper.

III. Types of Reordering Rules

In statistical machine translation, reordering can be done with the help of distance-based reordering model. This distance is computed relative to the previous phrase. Consider the notation to compute the reordering distance as:- $start_i - end_{i-1} - 1$. Where $start_i$ is the starting position of the source phrase that translates to ith of target phrase, end_i is the ending position of the source phrase that translates to ith of target phrase. The words skipped in translation process, is known as reordering distance. The Machine Translation is not capable to produce automatic, high quality and general purpose translation. A system is needed, that can check ambiguity and other issues of natural languages processing. The long sentences of English based on abstract concepts, such abstract concepts are not supported by Punjabi, so it leads to several problems in producing good translation. For high quality translation, the SMT system can handle both short distance and long distance sentences reordering. But, long distance sentence reordering required high time and space complexity. So, we need a generic rule that grammatically matches the source language sentences with the target language.

We classify reordering by using the width of the distance between the words that are moved during the translation. There are two classes of word width i.e. short distance and long distance reordering.

- Short Distance Reordering: An intuitive difference in word order between languages is the location of noun modifiers with respect to the noun. For example, adjective precede the noun in case of English language and in Punjabi it is followed by noun.

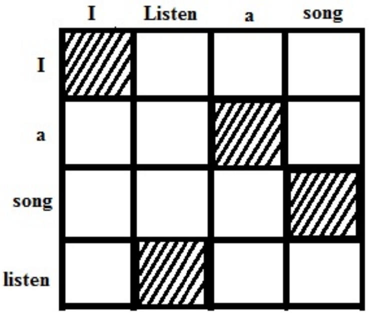


Fig. 1. Short distance reordering

As we see from the alignment matrix, to translate such a sentence, the decoder should allow a jump over or jump back the word or word group. This kind of permutation requires skip one or two words but less three.

- Long Distance Reordering: A long distance reordering where the languages has different word orders. For such long distance sentences, lexicalized distortion model is not sufficient for effective reordering.

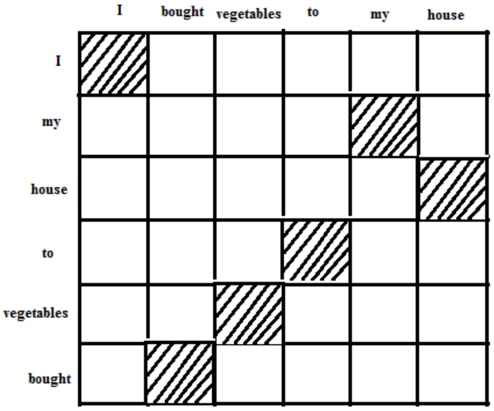


Fig. 2. Long distance reordering

IV. Reordering MEthodology

In preprocessing, reordering plays very important role for the sentence of source language. Reordering become significant when the syntactic structure of source and target language is different. The rules of reordering constructed manually according to the word order difference of both English and Punjabi. The sentence in a parallel corpus is reordered during the training process. Lexicalized automatic reordering is implemented in Mosestoolkit. Such automatic reordering is suitable only for short distance sentences.

The following example of English sentence shows the reordering before the actual translation.

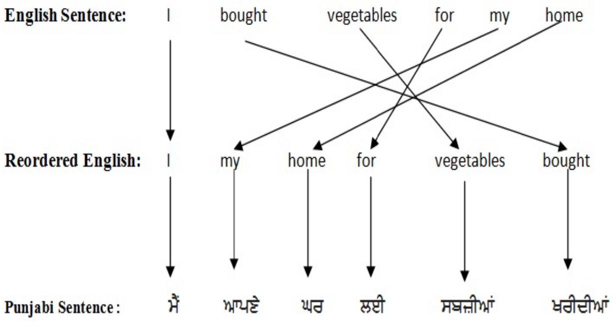


Fig. 3. Reordering source sentence

Fig. 4 shows the reordering of English sentence done by Stanford parser to retrieve the syntactic information. To obtain the word order according to target language (Punjabi), reordering is applied in source sentence before translation. Reordering rules are constructed to overcome the syntactic difference between English and Punjabi. All the rules included in Table 1.

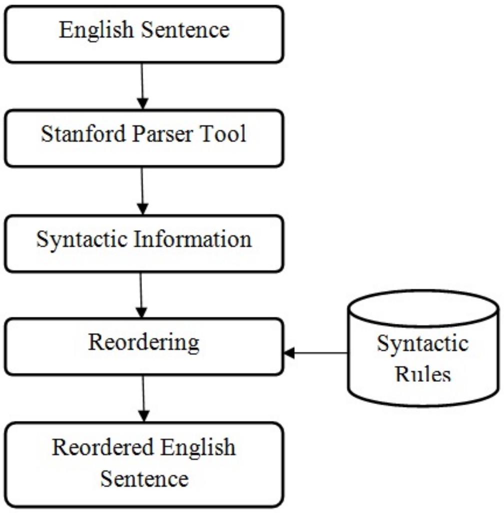


Fig. 4. Reordering process

There are three parts of Reordering rules:-

i. Production rules of source language(English).

ii. Transformed production rules according to Target language (Punjabi).

iii. The transformation of source sentence indicated by numbers.

Reordering rules are depicted in Table 1.

TABLE I. REORDERING RULES

Sr. No.	Source	Target	Transformation
1.	S->NP VP	#S->NP VP	#0:0,1:1
2.	PP-> TO NP PRP	#PP-> TO NP PRP	#0:0,1:1
3.	VP-> VB NP* SBAR	#VP-> NP* VB SBAR	#0:1,1:0, 2:2
4.	VP-> VBD NP	#VP-> NP VBD	#0:1,1:0
5.	VP-> VBD NP-TMP	#VP-> NP TMP VBD	#0:1,1:0
6.	VP-> VBP PP	#VP-> PP VBP	#0:1,1:0
7.	VP-> VBD NP NT-TMP	#VP-> NP NP-TMP VBD	#0:2,1:0, 2:1
8.	VP-> VBD NP PP	#VP-> PP NP VBD	#0:2,1:1, 2:0
9.	VP-> VBDS	#VP-> S VBD	#0:1,1:0
10.	VP-> VBS	#VP-> S VB	#0:1,1:0
11.	VP-> VB NP	#VP-> NP VB	#0:1,1:0
12.	PP-> TO NP	#PP-> NP TO	#0:1,1:0
13.	VP-> VBD PP	#VP-> PP VBD	#0:1,1:0

For instance, take a Sixth reordering rules from Table

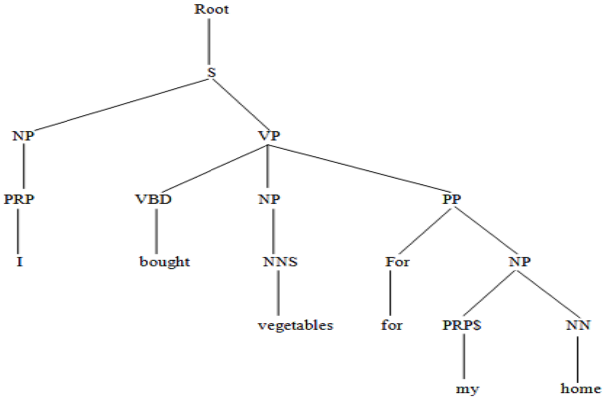
VP-> VBP PP #VP-> PP VBP #0:1,1:0

Where, # divides the units of reordering rules, this last unit indicates source and target indexes. In the above example, "0:1, 1:0" indicates first child of the target rule is from second child of the source rule; second child of the target rule is from first child of the source rule.

Source → 0(VBP) 1(PP)
Target → 1(PP) 0(VBP)

Production rules of English sentence are:-

i. S-> NP VP
ii. VP-> VBD NP PP
iii. PP->TO NP



iv. NP-> PRP $ NN

Fig. 5. English syntactic tree

Production rule i. S-> NP VP matched with the first reordering rule in Table 1. The Target transformation is same as the source pattern and therefore no change in first production rule.

Production rule ii. VP-> VBD NP PP is matched with the eighth reordering rule in table and the transformation is 0:2 1:1 2:0, it means that source word order (0,1,2) is transformed into (2,1,0),(0,1,2) are the index of VBD NP and PP, now the transformed pattern is PP NP VBD. This process is continuously applied to each of the production rules. Finally the transformed production rule is given below:-Reordered production rules of English sentence are as following:-

i. S-> NP VP
ii. VP-> PP NP VBD
iii. PP-> NP TO
iv. NP-> NN PRP $

Reordered English sentence: *I my home for vegetables bought*

English parallel corpus which is used for training is reordered.Majority of English sentences are reordered correctly by incorporating the above constructed rules. Original and reordered English sentences are shown in Table 2.

Table II ORIGINAL AND REORDERED SENTENCES

Original Sentences	Reordered Sentences
She visit last year	She last year visit
I listen a song	I a song listen
Ram gave his book to Sham	Ram his book Sham to gave
I eat bread.	I bread eat.
She loves him	She him loves

V. Conclusion

In this paper, we have introduced a new phrasal reordering model of integrating the phrase dependencies as syntactical structure to the Phrase-base SMT. We exploit the syntactically-informed reordering elements which are included by the translation direction feature in order to deal with the medium-and long-distance reordering problems. The proposed model has been discussed from the theoretical and experimental points of view, and its advantages, disadvantages and constraints in comparison of well-known and popular reordering models have been analyzed. In order to compare the performance of our reordering model with the distortion, lexicalized and hierarchical reordering models, lots of experiments have been carried out by training Persian, English SMT systems. We evaluated the proposed model on two translation tasks in different size. The evaluations illustrate significant improvements in BLEU, TER, and LRscore scores comparing to the lexicalized/distortion/hierarchical models. Furthermore, the reordering predictive capabilities of models have been compared by calculating the minimum number of shifts needed to change a system output so that it exactly matches a given references. The results imply that our model predicates a lot more reordering needed particularly medium- and long-distance reordering than the order reordering models. For a more detailed analysis and answering the question which word ranges are affected more by the reordering models, total precision/recall and precision/recall per distance have been calculated. The proposed model retrieved a significant impact on precision with comparable recall value respect to the lexical reordering model.

ACKNOWLEDGEMENT

Authors are highly thankful to the reviewers for their valuable suggestions and Department of Research & Development, IKG Punjab Technical University, Jalandhar for provide me facilities for the completion of this work.

References

[1] S. NieBen and H. Ney, “Morpho-syntatic analysis for reordering in statistical machine translation”, in the proceedings of MT Summit VIII, Santiago de Compostela, Galicia, Spain, pp. 247-252, 2001.

[2] F. J. Och and H. Ney, “Improved statistical alignment models”, *in the proceedings of the 38th Annual Meeting on the Association for Computational Linguistics*, pp. 440-447, 2003.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, R. Dyer, O. Bojar, H. A. Constantin, “Moses: Open source toolkit for statistical machine translation”, *in the proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2006.

[4] C. Tillman and T. Zhang, “A localized prediction model for statistical machine translation”, *in the proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*, pp. 557-564, 2005.

[5] C. Wang, C. Michael, and K. Philipp, “Chinese syntatic reordering for statistical machine translation”, *in the proceedings of the Empirical Methods in Natural Language Processing and Computational Natural Language Processing* (EMNLP-CoNLL), 2007.

[6] B. Chen, M. Cettolo, and M. Federico, “Reodering rules for phrase-based statistical machine translation”, *in the proceedings of the International Workshop on Spoken Language Translation Evalution Campaigen on Spoken Language Translation*, pp. 1-15, 2006.

[7] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation”, *in the proceedings of the Human Langauge Technology Conference* (HLT-NAACL), pp. 127-133, 2003.

[8] R. Zens and H. Ney, “A Compartive Study on reordering constraints in statistical machine translation”, *in the proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*, pp. 144-151, 2003.

[9] M. Galley and C. D. Manning, “A simple and Effective Hierarchical Phrase Recording Model”, *In the proceedings of the EMNLP*, 2008.

[10] Y. Gao, P. Koehn, and A. Birch, “Soft dependency constraints for reordering in hierarchical phrase-based translation”, *in the proceedings of the conference on Empirical Methods in Natural Language processing*, pp. 857-868, 2008.

[11] P. Koehn, J. Och, and D. Marcu, “Statistical Phrase-Based translation”, *in the proceedings of HLT/NAACL*, 2003.

[12] F. J. Och, “Minimum error rate training in statistical machine translation”, *in the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 160-167, 2003.