International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4

ISSN: 2454-4248
349 – 353

A Survey of Word Reordering Model in Statistical Machine Translation

Maitry Shukla
Research Scholar
Marwadi Education Foundation
Rajkot, India
shuklamaitri77@gmail.com

Prof. Harsh Mehta
Assistant Professor
Marwadi Education Foundation
Rajkot, India
harsh.mehta@marwadieducation.edu.in

Abstract— Machine translation is the process of translating one natural language in to another natural language by computers. In statistical machine translation word reordering is a big challenge between distant language pair. It is important factor for its quality and efficiency. Word reordering is major challenge For Indian languages who have big structural difference like English and Hindi language. This paper present description about statistical machine translation, reordering model and reordering types.

Keywords—*Machine translation,Statictical machine translation,word reordering, parser.*

*****

I. INTRODUCTION

Machine Translation (MT) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. Machine Translation system are needed to translate literary works which from any language into native languages[1]. The literary work is fed to the MT system and translation is done. Most of the information available is in English which is understood by only 3% of the population. This has led to digital divide in which only small section of society can understand the content presented in digital format. MT can help in this regard to overcome the digital divide. Statistical machine translation which is one type of machine translation is useful to build translator between the pair of languages which has similar word order. But is does not work well with distant language like English and Hindi, since English is an SVO language and Hindi is an SOV language. .English and Hindi language has different structure so for that it is difficult to get perfect translation. Word reordering is major task for those languages who have big difference in structure.

Word reordering is one of the most difficult aspects of statistical machine translation (SMT),and an important factor of its quality and efficiency.[8] Despite the vast amount of research published to date, the interest of the community in this problem has not decreased, and no single method appears to be strongly dominant across language pairs. Instead, the choice of the optimal approach for a new translation task still seems to be mostly driven by empirical trials. In this survey, we briefly explain different reordering models which are used in statistical machine translation. Problem of reordering is NP-hard itself. In this we explain distortion model, lexicalized model, tree-based model , chunk-based reordering.

This article has following sections, Section II contains the overview of Statistical machine translation and reordering problems, Section III discuss the literature survey of papers, Section IV provides the different models of word reordering in SMT, Section V describes types of reordering. We conclude our work in last section VI that is Conclusion.

II. STATISTICAL MACHINE TRANSLATION

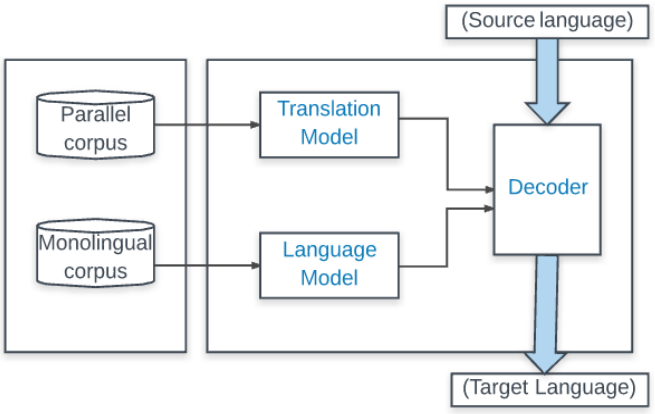

Fig. 1 SMT system

Statistical machine translation *(SMT) is a* machine translation *paradigm where* translations are generated on the basis of *statistical* models whose parameters are derived from the analysis of bilingual text corpora.

It gives the translation probability. It use Bay's theorem.[2] In that there are mainly three parts translation modeling, language modeling and decoder. Translation model gives the probability of each words translation and it use bilingual corpus. Language model use monolingual corpus and it rearranges the words as per the target language. Decoder will choose the translation which has highest priority. We suppose that the sentence *f* to be translated was initially conceived in language *E* as some sentence *e*. During communication *e* was

 ISSN: 2454-4248
349 – 353

corrupted by the channel to *f*. Now, we assume that each sentence in *E* is a translation of *f* with some probability, and the sentence that we choose as the translation (ˆ*e*) is the one that has the highest probability. In mathematical terms [Brown et al., 1990]

$$\hat{e} = \underset{e}{argmax} P(e|f) \qquad (2.1)$$

In above equation P(e|f) should depends on two factor.

1.The kind of sentences that are likely in the language *E*. It is known as the *language model* — *P*(*e*).

2. The way sentences in *E* get converted to sentences in *F*. It is called the *translation model* — *P*(*f*| *e*).

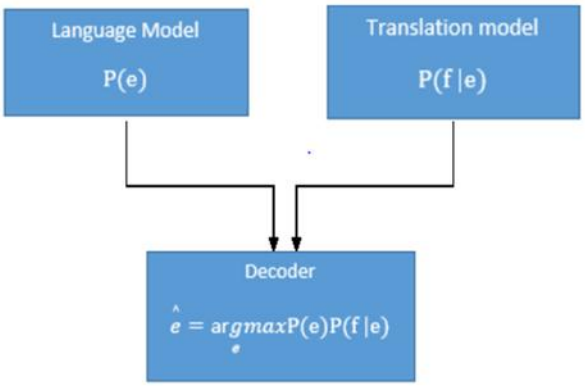

Fig. 2Noisy Channel Model for Machine Translation

In this it apply Baye's rule .Using Bayes' rule it break P(e|f) into two terms, P(e) and P(f|e) .

$$\hat{e} = \underset{e}{argmax} \frac{P(e)P(f|e)}{P(f)} \qquad (2.2)$$

Since F is observed as the sentence to be translated, P(F)=1

$$\hat{e} = \underset{e}{argmax} P(e)P(f|e) \qquad (2.3)$$

Translation models work by giving high probabilities to *P* (*f*|*e*) or *P* (*e*|*f*) when the words in *f* are generally translations of the words in *e[2]*.

In English and Hindi languages there are structural differences between them like English follows subject-verb-object (SVO)

And Hindi is a verb final language and it follows subject-object-verb (SOV) in a sentence. Some post-modifier in English became pre-modifier in Hindi. So these is the structure difference between two languages are most important during translation. Below example illustrate the SVO and SOV structure in these languages.In example S is the subject of thesentence, S_m is the subject modifier, and similarly for the verb (V) and the object(O).

Example

The president of America will visit the capital of Rajasthan.

(S) (S_m) (V) (O) (O_m)

amerikA ke rAXTrapati rAjasthAna kI rAjadhAnI kI sEra karenge

(S_m) (S) (O_m) (O) (V)

So we have to reorder our sentences to get the best translation for these type of languages.

III. Literature Survey

Marta R. Costa-jussà a, , José A.R. Fonollosa[3] , Latest trends in hybrid machine translation and its applications describe about hybrid machine translation which is combination of rule based and corpus based MT. It uses hybrid approach. Most of the research combines sources of information (rules and data), but there are also projects combining various corpus-based approaches. So for that it use hybrid type architecture which is combination of Rule based machine translation (RBMT) ,Statistical machine translation (SMT) ,Example based machine translation (EBMT). Good result produced by hybridization have some application like speech translation, cross-language information retrieval, computer-aided. its applications brings significant improvement because they allow the simultaneous exploitation of a variety of systems.

Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, Masaaki Nagata [4] , In this paper it use one method that separately translate the clauses in source sentence and reconstruct the target sentence using clause translation non terminals. It is used for long distance languages. It works on English to Japanese language. It use Divide and conquer, hierarchical phrase based SMT , clause segmentation , aligned corpus , clause translation model approach . In other approach like word reordering and preprocessing them face difficulty like when input phrases are long and require significant word reordering model. We can overcome by using clause segmentation based reordering. It gives good performance on long sentences translation. This proposed method provide more practical long distance reordering at the clause level. Better translation model training using fie aligned corpus. Due to some wrong inflection this divide and conquer produce dissiliency in final sentence translation. In future work It Can overcome limitation by integrating all clause translation but it require much larger space for decoding. It can be done for further.

Ondřej Bojar, Pavel Straňák, Daniel Zeman[5] , In this paper they briefly explain about the data issue which is arise due to English to Hindi machine translation. In Mt system adding some more parallel data may be caused by various problems such as different domains,bad alignment, noise in new data. They discuss several available parallel data sources and provide cross evolution result of their combinations. It also discuss about a tool for viewing aligned corpora which make it easier to detect difficult parts in the data. It give explanation about phrase based SMT. Give specification about different parallel corpora. Like Tides, Daniel pipes,Emile etc. it gives

 ISSN: 2454-4248
349 – 353

also normalization techniques. Tide and Emile both are independently acquired resources overlap. That combination give better performance on BLEU. Every parallel corpus has some errors. Problems of Hindi vocabulary so it is a big disadvantage.

Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, Pushpak Bhattacharyya [6] , This paper published in Association computational linguistic journal in 2014. They use NTCIR-9 (2000 test sentence) , NTCIR-8(1251 sen.) corpus . In this it explain the phrase based and factor based SMT for Hindi-English. Also show that the use of number, case and Tree Adjoining Grammar information as factors helps to improve English-Hindi translation, primarily by generating morphological inflections correctly. It shows preordering and post ordering methods. There is some missing words in parallel data so it translate some out-of-vocabulary data. They apply approaches like Super tag as factor , number case as factor , generating number , preprocessing chunk reordering , post processing , removing OOV in Proposed method. They proposed method for improving the quality of machine translation. It translate some out-of-vocabulary words which are present in phrase table but not present in lexical table .in so many system outputs are adequate but not fluent. So for fluency is use ranking approach. Factor base SMT achieve high score compare to baseline. Factor based SMT use in generating noun inflection in Hindi

Arianna Bisazza , Marcello Federico [7] , In This paper they give a brief description about word reordering Statistical machine translation. Word reordering is one of the most difficult aspects of statistical machine translation (SMT),and an important factor of its quality and efficiency. Despite the vast amount of research published to date, the interest of the community in this problem has not decreased, and no single method appears to be strongly dominant across language pairs. Instead, the choice of the optimal approach for a new translation task still seems to be mostly driven by empirical trials. To orient the reader in this vast and complex research area, we present a comprehensive survey of word reordering viewed as a statistical modeling challenge and as a natural language phenomenon. The survey describes in detail how word reordering is modeled within different string-based and tree-based SMT frameworks and as a stand-alone task, including systematic over views of the literature in advanced reordering modeling. They provided a comprehensive overview of how the word reordering problem is modeled within different string-based and tree-based SMT frameworks, and as a stand-alone task. To summarize, string-based SMT considers all permutations of the source sentence and relies on separate reordering models to score them. A growing part of the research community has converged on a positive answer to the former question, but the latter remains open to date. While the field keeps evolving around these questions, SMT has already reached the stage of applied language technology. In This they provided some practical knowledge to the developers. They also find a definitive solution to the problems about word reordering in SMT.

Piyush Dilip Dungarwal [8], In this paper they briefly explain some basic word reordering models which are used in Statistical machine translation. In this survey, They briefly explain various reordering models that are used with statistical translation models. Reordering model is one of the important component of any statistical machine translation system. Problem of reordering is NP-Hard itself. In this survey, they explain various reordering approaches that can be used to solve this problem. They explain simple distortion-based reordering which is used with phrase-based and factor-based models. They also discuss limitations of this distance-based approach. Then introduce a new source-reordering based approach to handle the reordering's based on structural information of the input text. It gives explanation about how to use parse trees and shallow parsing for source-side reordering.

IV. Different Models Of Reordering In Smt

Different reordering models are Distortion-based reordering, lexicalizedreordering, Tree-based reordering,and Chunk-based reordering.

A. Distortion-based Reordering

In early research on statistical machine translation, the reordering phenomenon, called distortion, was handled By distortion models.[2] These models attempt to estimate a probability distribution $^{d(i|j)}$ that computes the probability that the translation of the word in the i'th position in the source sentence appears in the j'th position in the target sentence. The first distortion model was proposed by Brown et al. in IBM translation models. In IBM models 2 and 3, distortion is modeled with a probability distribution D(j|i,ltarget,lsource), which predicts the target output word position j based on the source input word position i and the source and target sentences lengths. English sentence broken into I phrases and each English phrase translated into Hindi phrase.$Start_i$ is the starting position of the English phrase that translates to ith Hindi phrase.end_i is the ending position of the English phrase that translates to ith Hindi phrase. Now, reordering distance is computes as: $start_i - end_{i-1} - 1$.Reordering distance is nothing but the number of words skipped, while taking English words out of sequence.The major drawback of these models is that they do not generalize well since they assume that the reordering will occur in the same way for the words in the same position over different sentences, which is an unrealistic assumption Furthermore, these models do not consider the

 ISSN: 2454-4248
349 – 353

fact that in the translation, adjacent words do not move independently and tend to move together. In other words, adjacent words in the source sentence tend to stay adjacent in the target sentence as well. This limitation is overcome in IBM model 4 and5.

This model is preferable for those language pair who has similar word order.However, for the language pairs with different word ordering it is not a realistic one.For translation between languages with different syntactic structures long distance reordering is needed and since distortion model penalizes this reordering, it is not adequate for the translation between these types of languages.

B. Lexicalized reordering

In distortion models, reordering is only conditioned on the distances between the translated phrases and not the phrases themselves. However, it is observed that some phrases tend to be reordered with their adjacent phrases. Consequently, in the lexicalized reordering model, reordering is conditioned on the actual phrases.Koehn et al.
[7] propose the three reordering types 1)monotone 2)swap 3)discontinues If a phrase directly follows its previous phrases in the translation, the reordering is monotone, if it is swapped with a previous phrase, the reordering is swap, and if it is not adjacent with the previous phrase, the reordering is discontinuous.[8] It is learnt directly from the word alignment from data. In that if a word alignment point to the top left exists then it is a monotone reordering. If a word alignment point to top right exists then it is swap with previous reordering and If a word alignment point to top right exists and top left exists then it is discontinues reordering.

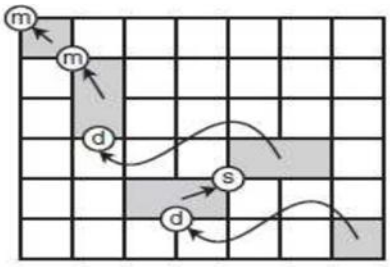

Fig 3. Example of lexicalized reordering

C. Tree based Reordering

For tree-based reordering we only consider repositioning the verbs at the end of the sentence or clause. We categorize each source (i.e., English) sentences into three basic types: simple, complex and compound, and reposition the verbs accordingly. For identifying the basic sentence type we first parse the source sentences. The tree outputs are categorized into the above mentioned three types by analyzing the structure of the tree and presence of key words such as that, which, who etc. and tags like CC, WHNP, SBAR, S.

D. Chunck based Reordering

While translating from source languages which don't have a constituency or dependency parser, it is very difficult to reorder the source sentence to match the word order as per the target language .[8]We can use shallow parsing techniques for source-side reordering. Chunk level tagging can be seen as intermediate annotations between POS tagging and parsing. The overall architecture of a translation system with chunk-based source reordering is shown in below fig.A reordering lattice is used for input to the translation system, instead of single sentence. Using lattice helps considering all possibilities of source-reordered input with their probabilistic scores. We first POS tag the input sentence and get chunk-level annotations. Then reordering rules are applied on these chunks to get a reordering lattice.

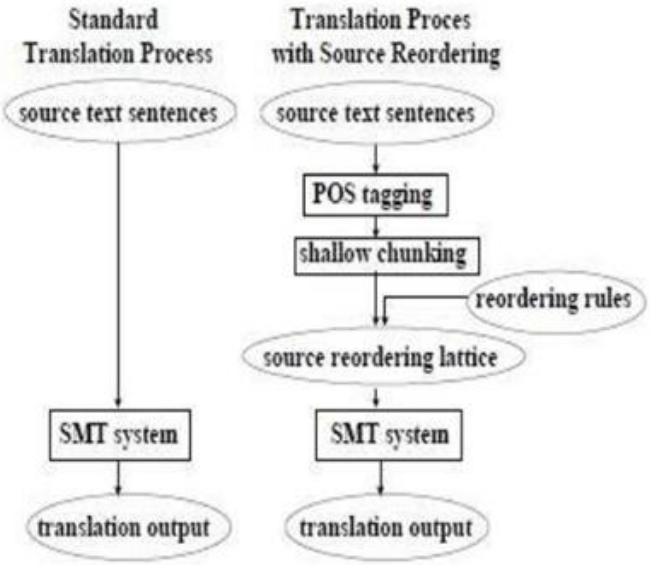

Fig.4Architecture of a translation system with and without chunk-based source reordering [Zhang et al., 2007]

V. TYPES OF REORDERING

There can be done in two way first is called pre-ordering and second is post-ordering.

A. Pre-ordering

While translating from source language to target language, to get best translation source words match with the syntax of target words. For that we reorder our source sentence to match the syntax of target languages.[7] Pre-ordering it means reorder the source sentence to match the syntax of target languages before translating into target language. source side reordering can be achieved by using syntactic parse tree of source sentences. In that either learn reordering rules and apply or need to find them manually. It helps to learn better word alignment and better phrase extraction. There are 5

 ISSN: 2454-4248
349 – 353

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).

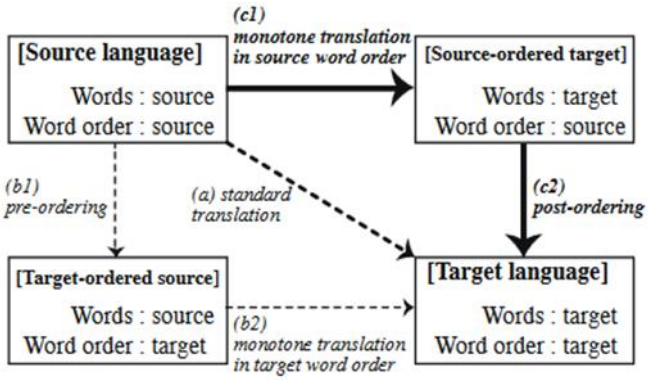

Fig.5Typical workflows of standard, pre-ordering, and post-ordering approaches to SMT. Taken from Sudoh et al. (2011)

B. Post-ordering

While translating from source language to target language, to get better translation we apply our reordering approach on target language.[7] In this type of reordering first it translate source language into the target language and then apply reordering approach on that translated output. So it reorder the target language to match the syntax of as per the target language. Above figure shows the post ordering process in that firstly source language translated into source word order now it has the translated words as a target language and then it post-order it and get better translation.

VI. CONCLUTION

This paper provided a comprehensive survey of different reordering model in Statistical machine translation. It describe types of word reordering. Also provide short description about statistical machine translation.

ACKNOWLEDGEMENT

The authors would like to thank editors, and reviewers for the valuable comments.

REFERENCES

[1] Koehn, Philipp. Statistical machine translation. Cambridge University Press , 2010.

[2] Peter F. Brown,Stephen A. Della Pietra,Vincent J. Della Petra, Robert L. Mercer» The Mathematics of machine Translation: Parameter Estimation"Journal Computational linguistics volume 19 issue 2, june 1993.

[3] Marta R. Costa-jussà a , José A.R. Fonollosa ," Latest trends in hybrid machine translation and its applications" 0885-2308/© 2014 The Authors. Published by Elsevier Ltd.

[4] Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, Masaaki Nagata , " Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation " Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics MATR, Association for Computational Linguistics, pages 418–427,Uppsala, Sweden, 15-16 July 2010.

[5] Bojar, Ondřej, Straňák, Pavel, and Zeman, Daniel."Data Issues in English-to-Hindi Machine Translation." In Proceedings of the Seventh International Language Resources and Evaluation (LREC'10), pages 1771–1777, Valletta, Malta, May.

[6] Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, Pushpak Bhattacharyya, "The IIT Bombay Hindi⇔English Translation System at WMT 2014 "Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics pages 90–96, Baltimore, Maryland USA, June 26–27, 2014.

[7] A Bisazza and M. Federico, "A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena", Computational Linguistics, vol. 42, no. 2, pp. 163-205, 2016.

[8] P. D. Dungarwal, "Reordering Models for Statistical Machine Translation: A Literature Survey," Indian Institute of Technology, Bombay, India, 2014