ISSN:0975-9646

Survey on Relative Clause based Text Simplification for Language Translation

Zenil Mahesh Prajapati, Bhanu Pratap Singh Bora, Nikhil Krishna Vhadge, Laxmikant Malphedwar
Department of Computer Engineering

Abstract-Language translation is research oriented area in today's world. . Info of grammar structure of source and target languages is must for interpreting one language to other. A compiler compiles a program written in a suitable source language into a corresponding target language through a number of phases. Preliminary with recognition of token through target code generation provide a basis for communication interface between a user and a processor in substantial amount of time. In the GLAP model different steps for generation of tokens through lexemes, and better input scheme implementation have been introduced. Disk access and state machine driven Lex are also replicated in the model towards its whole utility. Clauses are an essential part of any language and helps in making complex sentences in different contexts. This difficulty leads to a low score of translation in almost each and every Machine Translation engine existing. . In first phase input text is in English, save it in a file, extract words and punctuations from it and save them in an array, then clustering of words and find their meanings in context to the sentence and convert it to the target language (Hindi). Several issue occurred during the translation like word order, word sense, ambiguity and idioms.

INTRODUCTION

The process of translation of text from one language to another is known as Machine Translation (Machine Translation). Machine Translation system translates text from one natural language to another language. Machine Translation is needed to translate the text to our own native language or from to other usually known language. Language may cause a communication gap between people from various societies and cultures. NLP (Natural Language Processing) is the field of Computer Science that tries to fill this gap [1].

The first successful demo of Machine Translation system was done by the collaboration of Georgetown University and IBM in the year 1965. The importance of Machine Translation rises from the socio political importance of translation in societies where many language is spoken. The only possible others to rather widespread use of conversion is the adoption of a single common 'lingua franca'. A common language used by speakers of different tongues, which is not an attractive substitute for it, because it involves the authority of the chosen language, and to be disadvantage to speakers of other languages [2]. There are chances of other languages becoming ultimately disappearing except common language. As a language is related with a culture or a society, the loss of a language often leads to the vanishing of a distinctive culture; this is a loss that should matter to everybody. So translation is required for communication for day-to-day human communication and for gathering the information of various arenas. Everyone express his ideas best in his native language.

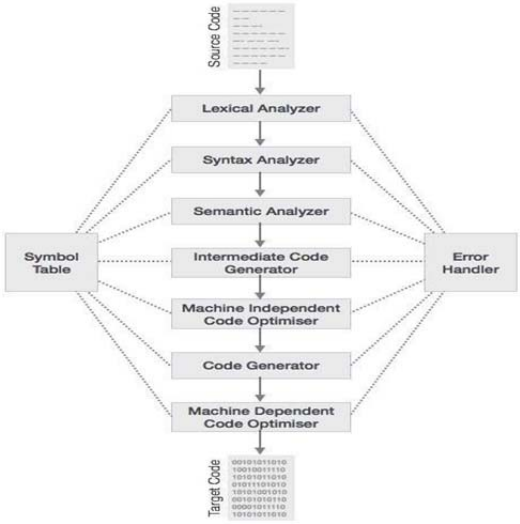

Fig 1: Flowchart to represent the different stages in a parser.

Machine translation also helps to overcome technical barriers. Most of this information is available in English which is understood only by a portion of population. To make this knowledge available by one and all, it has to be translated to many other national language of country respectively which are known to people. Example: A Germany scientist discovers a new theorem in a specific field which needs to be shared with all people through the world, but it can't be done as it is in Germany so it cannot be understood by everyone. But if it is translate to English which is known by popularly throughout the world then the problem can be solved.

Such statistical algorithms are performing very well in retrieving texts, splitting them into parts, checking the spelling and counting the number of words [3] . In most of the translation problems, where we have to deal with context based translation, these algorithms are not very accurate. All the existing word-based algorithms are limited by the fact that they can process only the information that they can see.

So, quality of machine translation system is one of the most important factors while designing any such system. A Machine translation engine can be evaluated on various aspects of its performance. Hampshire et.al. [4] have surveyed Free Online Machine Translators (FMT) systems and the results are shown in table I. The FMT Engines are evaluated and rated based upon 5 different performance parameters. These parameters are based on correct translations of idioms, registers, lexicons, phrases and grammar. A total performance score is used to rank the translation engine.

Translator	Rank	Total (Out of 25)	Idiom (5)	Register (5)	Lexic (5)	Phrasals (5)	Grammar (5)
Google	1	20	2	4	4	5	5
Babylon	2	16	5	1	4	5	0
Reverso	3	13	5	3	2	0	2
Bing	4	12	4	1	4	3	0
Babelfish	5	11	5	2	1	2	1
Systrans	6	9	5	2	1	2	0
Prompt	7	8	0	1	2	0	5
Wordlingo	8	4	0	1	1	2	0
Intertran	9	2	1	0	1	0	0
Webtrance	10	0	0	0	0	0	0

Table 1: COMPARISION OF THE PERFORMANCE OF FREE MACHINE TRANSLATION ENGINES

RELATED WORK

Text simplification can be a useful technique to achieve better translation. This process breaks the complexly constructed sentence into simple sentences by identifying the structure of the sentence. Lucia Specia [4] have worked on Translating from Complex to Simplified Sentences. In her work she has investigated the problem of simplifying Portuguese texts at the sentence level by treating it as a ”translation task”. In her approach she has used SMT framework to translate from complex to simplified sentences. The results of this work are promising according to BLEU evaluation technique. The proposed model is usually overcautious in producing simplifications. In the process the overall quality of the sentences is not degraded and certain types of simplification operations, mainly lexical, are appropriately captured.

Sanja et. al. [11] have worked on an automatic evaluation process for text simplification systems.

Their study explores the possibility of replacing the costly and time-consuming human evaluation in the process of text simplification systems. They have focused on six different widely used evaluation matrices for machine translation and try to find the corresponding correlation with human judgment.

In the process they are judging on the criteria of grammatical correctness and meaning preservation in text snippets. The work on sentence simplification is not only explored for English language but for all major languages worldwide.

Takao et. al. [12] have worked on Japanese sentence simplification task. They proposed a method to split and translate input sentences for speech translation in order to overcome the long sentence problem. Their approach have been built on three criteria used to judge the goodness of translation results which utilizes the output of an MT system only and assumes neither a particular language nor a particular MT approach. Aranzabe et al. [5] have developed a sentence simplification system for Spanish regional language Basque. In their work they focused on exact phenomena like appositions, fixed relative clauses and fixed temporal clauses. The simplification proposed does not eliminate any target people, and the simplification could be used for both machines and humans. Similarly Collados et al. have worked on Text simplification for Spanish language [5]. In India also this field is being well researched at different research centers. Hindi is the main language undertaken for this investigation. Rahul Sharma and Soma Paul at IIIT Hyderabad have worked on the clauses present in the complex sentences [6]. Their work presents, the task of identification and classification of clauses in Hindi text. Hindi is relatively unexplored language for text simplification and they have shown promising results in identification and classification for finite clauses.

Relative clauses are one of the most features clauses in constructing complex sentences.

Dornescu et al. [12] have specifically worked on Relative clause extraction for syntactic simplification. Their process reduced the average sentence length and complexity to make text simpler. The team analyzed the extraction of relative clauses through a tagging approach. A dataset covering three genres was physically annotated and used to develop and associate several approaches for automatically detecting appositions and non-restrictive comparative clauses.

Inspired by the work of Aranzabe et al. [8], Sharma et al. [11] and Dornescu et al. [12], we have targeted relative clauses present in the complex sentences and provide a better approach to identify and extract them for simplification process.

Existing Projects on Machine Translation-

a) **ANGLABHARATI-** This project is developed by IIT Kanpur in collaboration with the “ER&DCI”**.** In this project conversion of English to Hindi language was done using Rule based machine translation. In this project problem of ambiguity is solved. In this project all the possible values of ambiguous words is offered to the user, and then the user need to choose only the correct choice from it, and get the exact and correct meaning [1].

b) **MATRA-** In this project translation of English to Hindi language was done using Transfer Approach. This project used rule based approach to resolve ambiguities. As in Hindi the pre-positions of English becomes the post-positions so this basic concept is followed to solve word order issue.

c) **MANTRA-** This scheme is made from University of Pennsylvania. In this project translation of text from English to Hindi and also focused on input text. When the input is entered then it checks whether it is in fine grammatical form or not, if not then it should be firstly corrected and then translated [3].

d) UCSG MAT- Machine Aided Translation project was proposed by University of Hyderabad to translation of English to Kannada.

PROPOSED WORK

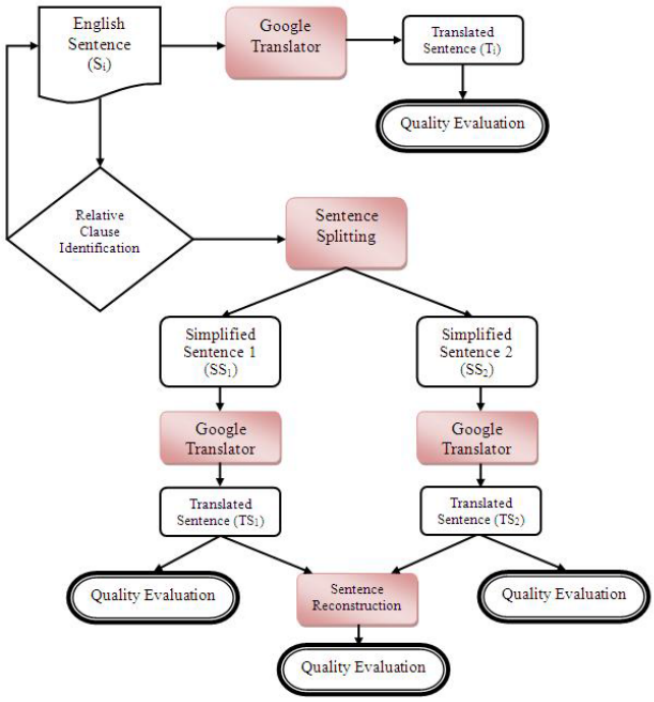

Fig 2: Flowchart for the proposed method to improve the quality of translation for English to Hindi text

We have focused on English to Hindi translation score improvement in this proposed work. Maximum Entropy Inspired Probabilistic LFG F-Structure Parsing techniques are used to parse the sentence tree. The proposed method for improving the translation score of English to Hindi texts is explained through flowchart shown in Fig 2. The whole process comprises of following steps:

Relative Clause Identification Presence of relative clause in a sentence can be identified by a relative pronoun at the start of the clause. Although sometimes we can infer this simply by word order. The choice of relative pronoun, or choice to omit one, can be affected by two factors, Human and Non-human. There are two types of relative clauses, Restrictive and Non-restrictive. These can be distinguished by the presence of pause in the speech or comma in the text.

Splitting: In this step we parse the source tree and identify the location of relative clauses in the structure. This step is also the core of the proposed method. We have parsed the given tree with Maximum Entropy Inspired Probabilistic LFG F-Structure Parser [11]. This parser is phrase based parser and suitably modified for identification of different phrases present in the parse tree. With the help of this parser, we obtain the grammar tokens for each word and phrases present in the tree. The parser labels words with part-of-speech tags, such as ".n" (noun), " .v" (verb), " .a" (adjective) and " .e" (adverb).

If we consider a sample input sentence"A notebook is a computer which can be carried around", then the corresponding parse tree for the sentence is shown in Fig 2.

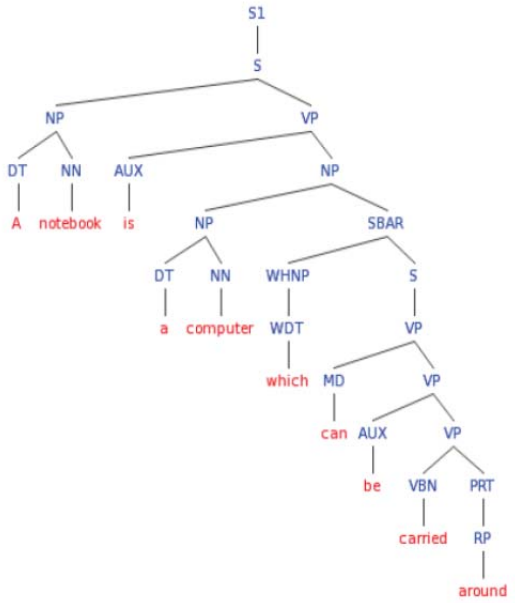

Fig 2: Parse tree generate for the sentence "A notebook is a computer which can be carried around".

Machine Translation In the next stage, simplified sentences *SSis* are given as input to the Free Machine Translation Engines (FMT). The FMT is embedded in the system and with *SSi* as input sentences, target translation is produced. The outputs from the FMT, *TSi*, are the translated simplified sentences. In our method we have used Google

Translator1 for English to Hindi translation because it is one of the state-of-the-art free machine translation systems available today with best BLEU score. Although Yahoo Babel Fish and Windows Live Translate 3 are also available for free use, but the evaluation matrices in Table I suggests that Google Translator is best among these. Fig 3 shows the results of translations obtained using Google Translator for the sample sentence and corresponding split sentences.

English Sentence	Hindi Translated Output	Quality Score of translation
A notebook is a computer which can be carried around (S_i)	एक नोटबुक चारों ओर ले जा सकता है, जो एक कंप्यूटर है। (T_i)	3
A notebook is a computer (SS_1)	एक नोटबुक एक कंप्यूटर है (TS_1)	5
A notebook can be carried around (SS_2)	एक नोटबुक चारों ओर ले जा सकते है (TS_2)	4

Fig 3: Text simplification results and quality scores for sample sentence "A notebook is a computer which can be carried around"

Evaluation: This is introductory work on the sentence simplification and we have tested algorithm on 2000 sentences. Since any machine evaluation system would not be properly trained and provide satisfactory results for small data sets, thus we have used manual evaluation process to evaluate the quality of translated sentences *TSis*. A sample sentence translation quality evaluation is shown in Fig 3. This is verified from the results that simplified sentences have better translation quality score as compared to the complex ones.

CONCLUSION

Machine translation is a process that helps in translating of the given text form to the required target language. Thus, Machine translation helps in overcoming lingual barriers and technological barriers. Also an MT system can replace a human translator which helps in minimizing our precious time and expenses too. English to Hindi Machine Translator System for viewers has been developed. By using example based approach trained this system and takes several sample texts in English and their corresponding Hindi sentences. Database is used to store text in source and target languages which is referred during translation process. Also maintained dictionary and encyclopedia to store words which have no meaning in Hindi and grouped them under former class and the latter stores the remaining.

REFERENCES

[1] Rashmi Gupta, Nisheeth Joshi and Iti Mathur, "Analyzing Quality of English-Hindi Machine Translation Engine Outputs Using Bayesian Classification", Apaji Institute, Banasthali University, Rajasthan, India, July 2013.

[2] Gennadi Lembersky,Noam Ordan, ShulyWintner, "Language Models for Machine Translation: Original vs. Translated Texts", Association for Computational Linguistics,2012

[3] Saini, Sandeep, and Vineet Sahula. "A Survey of Machine Translation Techniques and Systems for Indian Languages." Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on. IEEE, 2015.

[4] Hampshire, Stephen, and Carmen Porta Salvia. "Traslation and the Internet: Evaluating the quality of free online machine translators." Quaderns: revista de traducci 17 (2010): 197-209.

[5] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[6] Doddington, George. "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics." Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., 2002.

[7] Klakow, Dietrich, and Jochen Peters. "Testing the correlation of word error rate and perplexity." Speech Communication 38.1 (2002): 19-28.

[8] Lavie, A., and Satanjeev Banerjee. "The METEOR Automatic Machine Translation Evaluation System." (2005).

[9] Aranzabe, Mara Jesus, Arantza Daz de Ilarraza, and Itziar Gonzalez-Dios. "First approach to automatic text simplification in Basque." Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012). 2012.

[10] Specia, Lucia. " Translating from complex to simplified sentences." In Computational Processing of the Portuguese Language, pp. 30-39. Springer Berlin Heidelberg, 2010.

[11] Stajner, Sanja, Ruslan Mitkov, and Horacio Saggion. "One Step Closer to Automatic Evaluation of Text Simplification Systems." In Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pp. 1-10. 2014.

[12] Dornescu, Iustin, Richard Evans, and Constantin Orasan. "Relative clause extraction for syntactic simplification." COLING 2014 (2014)