Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 2534–2544

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Corpus based Machine Translation System with Deep Neural Network for Sanskrit to Hindi Translation

Muskaan Singh[a],Ravinder Kumar[a],Inderveer Chana[b]*

[a]*Language Engineering and Machine Learning Research Labs, Thapar Institute of Engineering and Technology,Patiala,India.*
CSED, Thapar Institute of Engineering and Technology, Patiala, India.

Abstract

Sanskrit language is the mother of almost all Indian languages. The main requirement in Sanskrit domain is to translate the life-transforming stories (epics), Vedas etc. to make them available in other languages, for public at large. In the field of machine translation system there is need to develop a machine translation system which translate Sanskrit language to Hindi. So, the main focus of this work is to propose a new corpus-based translation system for Sanskrit to Hindi translation where Bhagvad Gita – the song of the lord is used as an input data. In this work, Deep neural network is used for training where input data is passed to neural network after data analysis and processing which then performs auto-tuning that help to make this model better. Target text is prepared using this proposed model and achieves better BLEU Score and Word Error Rate.

Keywords: Corpus, MTS, Sanskrit, Hindi, Deep Neural Network.

1. Introduction

The translation of one language to the other language is the only aim of the Language translation Systems. This translation helps in socialization means people can communicate with each other in an easy way. As human is a

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: muskaan_singh@thapar.edu

1877-0509
10.1016/j.procs.2020.03.306

social entity, and from the ancient time they love to communicate with each other and live together. They communicate with each other by using different means of communication and exchange their data/information/thoughts with each other. Initially Sign Language is also the means of communication and is used to exchange thoughts with each other but now in this era of technology there are number of effective sources that are used for communication purpose. There are more than thousands of languages used all over the world even in India due to its diverse nature and culture, different languages are used. In the local areas of the India, they have used their own language of communication and Most of the literature is also available in the local languages. In India, there are number of languages are used for the purpose of communication. Sanskrit is the mother of almost all Indian languages and almost every language is generated from Sanskrit Language. As Sanskrit is very ancient language so the main requirements in Sanskrit domain is to translate the life-transforming stories (epics), Vedas etc. to make them available in other languages, for public at large. It is observed from various surveys that Sanskrit as a source or target language is in developing stages and a major issue which arises in implementation of Sanskrit based MTS is the approach used for developing it. There are number of historical and ancient granths which are written in Sanskrit language, Bhagvad Gita is one of them and is very valuable in Indian Hindu culture.

Bhagvad Gita (BG) – the song of the lord included in Bhishma Parva (chapter 23 - 40) which is sixth book out of 18 books of Indian Epic Mahabharata [29]. There are 18 chapters and 700 slokas (verses) in BG. In BG, there are number of combinations of different Hindu thoughts in regards to mystical bhakti, dharma and different yogic path to find moksha. It is an essential content which abridges the Upanishadic lessons and is remarked upon and translated by different schools of Indian theories. Being a vital content, it has been converted into every single real dialect of the world, and furthermore remarked variously. In this way annotators in uncertainty can generally allude to these critiques for right explanation. This sacred text being lucid and finish in itself, can be utilized for larger amount examination, for example, talk investigation, point identification, anaphora goals, etc. So, it is an important issue to convert this content and use it not only for research purpose but to save the history and culture of India. To make this conversion automated, the field of Natural Language Processing (NLP) may be used. A lot of researchers also worked on this to automate the translation systems and most of them selects NLP for the development of the translators for the conversion of one language to other. NLP is defined as a mechanism which helps in the understanding of the semantics and grammar of a language and accurate interpretation of the language [3]. NLP is concerned with the development of models that automate work to process a language and help to make its use for communication between humans. The basic model of the language processing system as shown in fig 1, requires some knowledge to deal with the processing of the language and it also require some knowledge about language along with its grammar [3]. Machine translation systems (MTS) are the systems which are totally based on NLP. In today's life, Machine translation is one of the applications of computers used to translate one natural language text into the other language and it is an application of NLP. The term Machine Translation System called as MTS also defined as the "translation from one natural language (Source Language -SL) to another language (Target Language -TL) using computerized systems with or without human assistance". These systems provide the translation solution without any human interference or assistance. The main requirement of these systems is dictionaries and grammar for those particular languages.

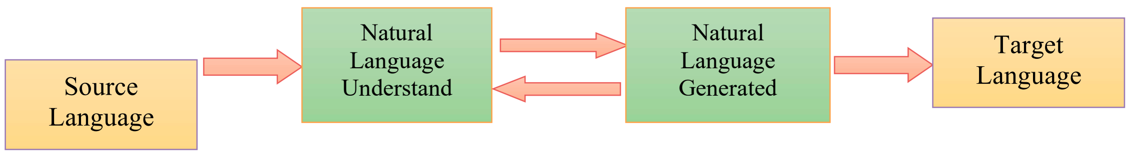


Fig 1: NLP Processing

1.1 Overview of MTS

With the advancement in the technology, the use of NLP is increased for the solution of different tasks. Now-a-days

Machine translation is one of the important tasks of NLPwhich is on high demand with the qualitative and efficient results. These qualitative results benefitted into various fields [3] in various terms as discussed below:

- ***Searching and Extraction of Data:*** In the Era of technology, Internet becomes the source of information for any field and for any one. So, the webpages where content is available might contains different languages so, translation is required to enable the search and to extract data and remove language barriers.
- ***Translation of Technical Documents:*** The technical documents of different types like patent, research articles, manuals and others professional documents which are used in various countries need to be translated so everyone can have access to that data.
- ***Translation of Documents and Speeches:***Some of documents and speeches which needs to float in multiple countries need translation in multiple languages. For instance, In European Union (EU) they used 23 different official language because each member of European parliament speaks his/her own language, because of this translation is required for each of these languages. The other important example is the translation of the documents of United Nation is 6 different languages.
- ***Translation of Broadcasted Information:***the media is one of the important sources of information where information is broadcasted continuously and this public information is flowing all over the world so, translation is required. There are many platforms that deals with the information transfer and these are Radios, blogs, Newsfeeds, Television, and many more.
- ***And many more.***

As discussed that the qualitative and efficient MTSs becomes the necessity of today, but the current technology which develops the MTSs is not up-to-the-mark means not fulfil the expectations to deal with all the defined areas but its continuous improvement originates the automated translation systems that are used for extracting information or data.

1.2 Approaches used in MTS

In MTS, Source Language is first analyses and its internal representation is prepared accordingly and then it is manipulated for transformation of it into its target language from which target language is generated. These transformations don't change the meaning of the input text throughout the whole process while translation. It may be bilingual or multilingual. If translation is between the two language then it is known as bilingual systems but it is between number of language then it is called as multilingual system [21]. The process of MTS may be defined in two steps:

Step 1: Decoding: where meaning of the source text is decoded
Step 2: Re-encoding: where decoded text is re-encoded in target language.

Decoding is the process where source text is analyzed along with its all features so that translator will get the accurate meaning of the source text. This analysis requires deep knowledge of the source language and other parameters like, grammar, semantics, syntax etc. and similarly for re-encoding.

There are Number of MTSs which developed for the translations of some common languages like, English, Chinese, Russian, Spanish, Japanese, Hindi and many other Indian and Foreign Languages. To develop these systems, different approaches of MTSs has been followed and these are: (a) Rule Based Approach, (b) Dictionary based Approach, (c) Corpus based Approach, (d) Knowledge Based Approach and (e) Hybrid Approaches as described in Fig2.

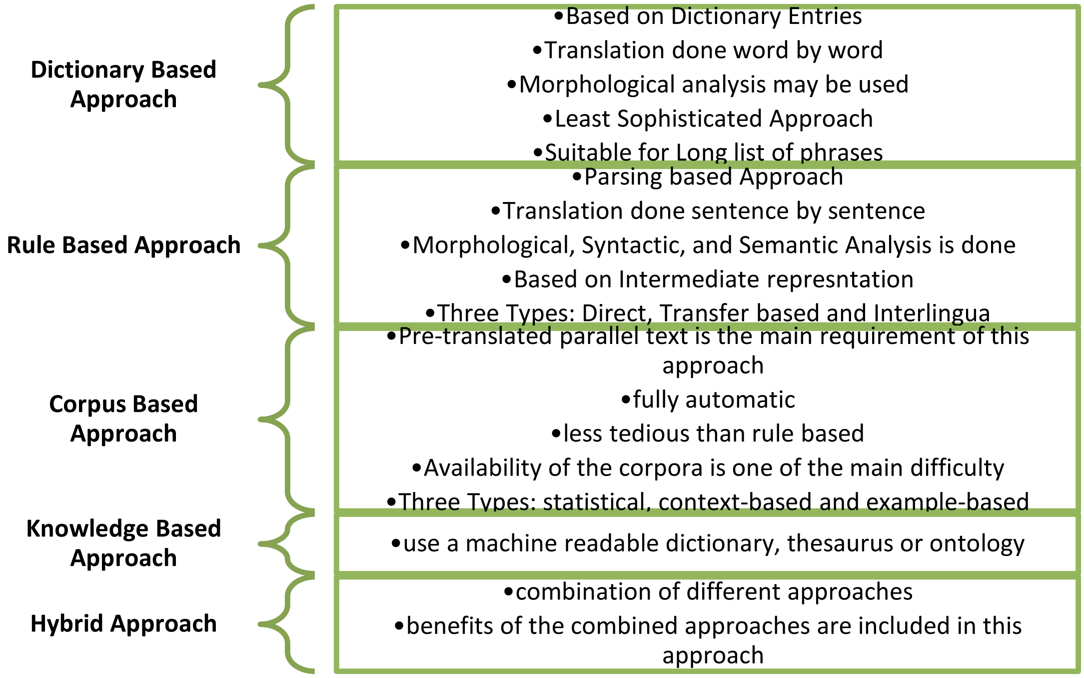


Fig 2: Approaches of MTS

1.3 Proposed Work

A new corpus-based machine translation system for Sanskrit to Hindi translation is proposed. There are very few works done on the translators which translated Sanskrit to Hindi Language so need to develop an efficient translator for the same. India is rich in culture and there are many Vedas and other holy books which were written in Sanskrit like, Bhagavad Gita. For this work, Bhagavad Gita(BG) is selected as an Input Data Set and proposed a translator which automatically translate it into Hindi Language.

The next sections of this paper describe some of the existing MTS for Sanskrit Language and the detail discussion of this proposed MTS along with its simulation and results.

2. Existing Sanskrit based Translation System

Different researchers done work on Machine Translation Systems for Sanskrit Language. The discussion on some of the proposed systems is given in this section. Tapaswi et al. [1] proposed a parsing technique named as Lexical Functional Grammar (LFG) for Sanskrit text. LFG works on two basic types of syntactic representation: (a) constituent structure, and (b) functional structure. They used LFG because translation was from Sanskrit to English and both these language representations is different. For instance, English is Subject-Verb-Object (SVO) and Sanskrit is Subject-Object-Verb (SOV). Their main aim to develop a parsing technique and testing of this is done on simple sentences. The other work on lexical analysis is done by Tapaswi and Jain [2] for Sanskrit sentences. In this work, morphological analysis was also added with lexical analysis. They designed their own rule format and stored all the rules in the files with the names of its starting letter like 'himalaya' is stored in 'h.txt'. Here the root word and meaning of it is identified with the help of lexical analyzer.

Barkade et al. [3] proposed MTS for the translation of English to Sanskrit Language where they divided work into 4 modules: Lexical Parser, Semantic Mapper, ITranslator, and Composer and discussed first two modules in this paper. They designed their own lexical parser for POS tag information and its dependency. Three different rules are generated by this parser and are: Equality Rule, Synonym Rule, and Antonym Rule. After parsing when tokens were

generated and using dependency when relation between token is found then tree was generated and mapping done between English and Sanskrit Sentence.EtranS- MTS for the translation of English to Sanskrit Language proposed by Bahadur et al. [4] to improve quality of translation. They developed a software using .Net Framework and MS-Access 2007. This software has two modules: (a) Parse Module where parsing was done means after analysis tokens were generated and grammatical and syntax analysis was done. (b) Generator Module uses semantic information for mapping and on the basis of mapping results were generated. They tested their proposed system for small, large and extra-large sentences and achieved 99% correctness for small and 90% for extra large sentences.
Mishra and Mishra [11] compared different MTS systems and analyzed the performance of Example based system for English to Sanskrit translation. They used ENGCC parser for English and for Sanskrit parser is used which was developed by Gerard Huet(http://sanskrit.inria.fr/) and dictionary is used from(www.dicts.info/dictionary.php?l1=English&l2=Sanskrit) site. They analyzed that the EBMT is used for scarce online resource's data and worked properly for results. The other work done by Mishra [24] where they proposed English to Sanskrit based MTS using rule-based approach. Pandey and Jha [18] analyses error for Sanskrit to hindi MTS that uses statistical approach. They build corpus and trained using MTHub platform. The error report generated by MTHub System and during training of two phases BLEU score was calculated. In the first phase, 10000 long, complex and compound sentences were used where BLEU score was 39.17 and in second phase, 24000 bilingual and 25000 monolingual sentences were used and achieved 41.17 BLEU score.

3. Proposed Corpus Based Machine Translation System with Deep Neural Network

The main focus of this work is to propose a novel approach to develop corpus-based MTS with deep neural network for translating Sanskrit to Hindi Text and here corpus-based MTS is developed which doesn't need any rules or dictionaries because they automatically learn about language from a large set of corpora. The corpus is improved with the addition of trained data using deep neural network to learn all the expressions like phrasal and idiomatic expressions easily. This work is done in three different phases named as (a) Data Analysis, (b) Data Processing and (c) Target Generation where firstly data analysis is done and then data will transfer and generated into target language and lastly final generation of target text using Deep Neural network. Here, deep learning provides better training and accurately find the results.

Proposed Model

This proposed corpus-based model with deep neural network divides the work into three phases named as: (i) Data Analysis, (ii) Data Processing and (iii) Target Generation. These phases further perform different operations to generate target text as shown in figure below:

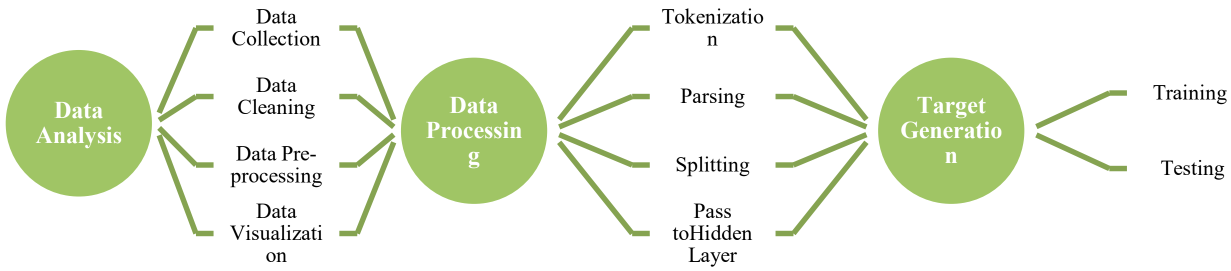


Fig 3: Phases in Proposed MTS

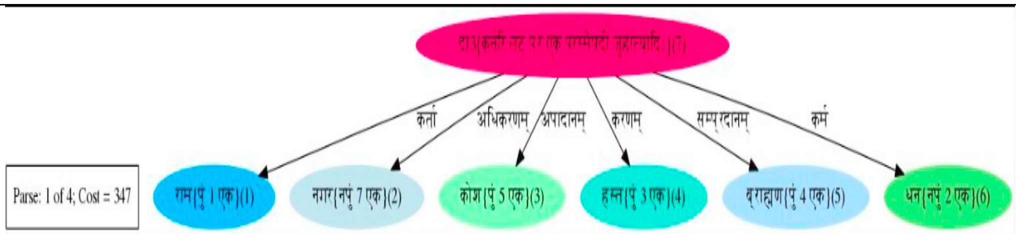
Phase 1: Data Analysis

This step includes the collection the dataset containing millions of word meanings in a systematic way from the different resources to make it useful for proposed model. From all sparse spreader datasheets make big dataset and save it in a proper word meaning format. If dataset is not in correct format, mark it as redundant dataset. After collecting all the dataset in a single csv, clean the dataset by removing all the redundant or unused words and extra symbols present in dataset. Here the dataset is arranged in a proper structure to make it suitable for the proposed model. After that, visualize the dataset to check the structure and the co-relation in the dataset. If there is no relation found in dataset then pre-processing of the data is again performed from starting.

Phase 2: Data Processing

Tokenization is the chopping of data into words. In this step store all the dataset in matrix format and then Tokenize the data by splitting all the sentences into words and label the data into numeric form. As this data is in Sanskrit and Hindi language, encode it normally as done for English words So here use Count vectorise. Count-Vectorise will convert all type of data to numeric form easily so use it for labelling the data. After tokenization data is analysed through lexical and semantic analysis. Here arrange the data in a proper grammatical structure with respect to the Sanskrit rules. To overcome the problem of overfitting and under fitting, the data is split in two parts out of which one is used for the model development for predictive analysis and the other one is used for performance analysis. Divide the data into two parts for training and testing purpose.

Table 1: Example of Data Processing

Input Sentence	<s>rAmaH nagare koSAw haswena brAhmaNAya XanaM xaxAwi.</s
Tokenization	rAmaH nagare koSAw haswena brAhmaN. XanaM xaxAwi
Semantic Analysis	1 rAmaH karwA,7 2 nagare aXikaraNam,7 3 koSAw apAxAnam,7 4 haswena karaNam,7 5 brAhmaNAya sampraxAnam,7 6 XanaM karma,7 7 xaxAwi
Parsing	
Morphology	6 2 13 7 1 #karma 1.20 6 1 7 7 1 #karwA 1.2 5 1 19 7 1 #sampraxAnam 3.5 4 1 18 7 1 #karaNam 3.1 3 1 20 7 1 #apAxAnam 3.9 3 1 26 7 1 #hewuH 3.10 2 4 24 7 1 #aXikaraNam 3.17 2 3 52 7 1 #samboXyaH 3.18 2 2 13 7 1 #karma 1.20 1 1 7 7 1 #karwA 1.2
Target Sentence	1.1 राम नगर में कोश से हाथ से ब्राह्मण को धन को देता है

Phase 3: Target Text Generation

In the proposed system, training is performed using deep neural network. Here input is passed neural network with a huge number of hidden layers that layers will be added automatically through auto-tuning. The auto-tuning has helped to make this model better than all other previous models. Now, system will automatically calculate required number of input layers which makes proposed system accurate and then it is trained using large number of hidden layers at a very high speed which makes it much better.

So after processing, now pass the training set into model. Here, Input is passed through first layer of neural networks and number of neurons. Secondly, applied activation functions and gives probability of output. At the end, it passes output of first NN to second layer NN. Input layer is the layer which interacts with hidden layers. The accuracy of the system is depended on the number of input layers passed and the number of times it interacts with the hidden layers. More the interaction between the input layers and hidden layers more is the accuracy attained and vice versa. Now, first of all, normalize the batch and again apply activation function. Passes output of second NN to third layer NN. Again, normalizing the batch and apply activation function. Then Passes output of Third NN to fourth layer NN. Repeat this process up to seventh layer NN. After addition of layers, divide and merge the datasets according as per the requirement. Normalize the batch again. Again, applied activation functions and gives probability of output. Again, divide and merge the data as per requirement neurons are passed in dense step. Then output is generated at the end that exhibits the accuracy of 99.97%.

4. Experiment Analysis

In this section, details related to experimental analysis has been discussed for corpus-based machine translation system with deep neural network.

4.1 Simulation Setup

In the proposed work, Keras sequential model is used to process the data. Proposed model is processed through highly configured core GPU with 32 GB of RAM to achieve a high throughput speed approximately 2500 words per second. This speed is not possible for normal systems because in this one epoch will take approximately two hours to run. Sothe use of highly configured GPU along with NVIDIA Geforce GTX 980 GPU is preferred here.

4.2 Dataset Used

For the proposed model huge amount of dataset is collected which consists of Sanskrit to Hindi word meanings as shown in Table 2. Along with the word meanings data it also consists of various rules of Sanskrit as well as Hindi. Thus, redundant data from that dataset has been removed and arranged in proper word meaning format to pass it through the model. Also, collect the huge dataset in a single csv file which contains around millions of sentences.

Table 2: Dataset

Dataset	#Sentences	#Words		Vocabulary	
		Source	Target	Source	Target
Corpus: Open MT Sanskrit-English					
Training	145,34,215	131575835	123425654	355.465	124.278
Development	192679	172799	122645	NA	NA
Test	12698	77322	26273	NA	NA
Corpus: Open BTEC English-Hindi					
Training	123643124	11425429	11753927	428.672	111.249
Development	12679	122769	87286	NA	NA
Test	10684	67827	39207	NA	NA

The collected data was non-uniform and unformatted data. Therefore, various ways are needed to make it suitable for proposed model. Single csv file is processed twice for rechecking to remove redundant words. If any redundant words are found it removes them from there, otherwise it does encode the dataset.

4.3 Corpus Based Results

BG has 700 sections. Greater part of these refrains (645) are created in a measurement called anustup and the rest of the sections are in indravajra, upendravajra and upajati meters. A series of characters isolated by spaces may compare to at least one words. The aggregate number of string successions isolated by spaces is 6426 adding up to 9.18 words per section. In the wake of part these strings into words, there were 8884 words 13 at the end of the day after division there was around 13.82% expansion in the words. Out of 8884, 1413 words were observed to be mixes adding up to 15.9%. The some of the resultant Hindi text generated by using this proposed Model is as shown in table 3.

Table 3: Sanskrit Sentences and their corresponding Hindi Sentences generated by using proposed model

	Sanskrit Sentences	Hindi Sentences
1	xk.Mhoa L=alrsgLrkÙoDpSoifjng~;rsA u p ‘kd~uksE;oLFkkrqaHkzerho p esaeu%AA	gkFk ls x.Mho /kuq’kfxjjgkgSvkSjRopkHkhcgqrtjjgh g SrFkkesjkeuHkzfer&lkgksjgkgS bl fy;seSa [kMkjgusdksleFkZ ugh gqaA
2	fufeÙkkfu p i’;kfefoijhrkfuds’koA u pJs;ks·uqi’;kfegRokLotuekgosAA	gsds’ko! eSy{k.kksa dk Hkhfoifjrgh ns[k jgkgqWrFkk ;q) essaLotuleqnk; dksekjdjdY;k.kHkh ugh ns[krkA
3	udkM~{ksfot;ad`”.k u p jkT;alq[kkfupA fdauksjkT;suxksfoUnfdaHkksxSthZforsu okAA	gsd`”.k! eS u rksfot; pkgrkgqWvkSj u jkT; lq[kksadksghAgsxksfoUn! gesa ,slsjkT; ls D;kiz;kstugSvFkok ,slsHkksxks ls vkSj thou ls HkhD;kykHkgSA
4	;s”kkeFksZdkM~+f{krauksjkT;aHkksxk% lq[kkfupA rbes·ofLFkrk ;q)s izk.kkaLR;DRok /kukfupAA	gesaftudsfy;sjkT;]HkksxvkSjlq[kkfnvHkh’VgSa] osgh ;s lc /kuvkSj thou dh vk’kkdksR;kxdj ;q) es [kMsgSaA
5	vkpk;kZ% firj% iq=kLrFkSo p firkegk%A ekrqyk% Üo’kqjk% ikS=k% ‘;kyk% lEcfU/kuLrFkkAA	xq:turkm&pkps] yMdsvkSjmlhizdkjnkns] ekes] llqj] iks+= lkysrFkkvkSjHkh lEcU/khyksxgSA
6	,rkUugUrqfePNkfe ?urks·fi e/kqlwnuA vfi =SyksD;kjkT;L; gsrks% fdauqeghd`rsAA	gs e/kqlwnu !eq>s ekusijHkhvFkokrhuksayksdksa ds jkT; ds fy;sHkheSabudksekjukughpkgrk! gQji`Foh ds fy;srksdgukghD;kgSA
7	fugR; /kkrZjk”VªkUu% dkizhfr% L;kTtuknZuA ikiesokJ;snLekUgRoSrkukrrkf;u%AA	gstuknZu ! /k`rjk”Viq+=ksadksekjdjgesaD;kizlUurkgksxh ! buvkrrkf;;ksadksekjdjrksgesaikighyxsxkA
8	rLekUukgkZo;agUrqa /kkrZjk”VªkULockU/koku~A LotuafgdFkagRoklqf[ku% L;keek/koAA	vr ,ogsek/ko ! viusghckU/ko /k`rjk”Vª ds iq+=ksadksekjus ds fy;sge ;ksX; ugh gS%] D;ksafdviusghdqVqEcdksekjdjgedSlslq[khgksxsaA
9	;|I;srs u i’;fUrykSHkksigrpsrl%A dqy{k;d`ranks”kafe=nzksgs p ikrde~AA dFka u Ks;eLekfHk% ikiknLekfUuofrZrqe~A dqy{k;d`ranks”kaizi’;fHntZuknZuAA	;|fi yksHk ls Hkz”Vfprgq, ;s yksxdqy ds uk’k ls mRiUunks”kdksvkSjfe+=ksa ds fojks/k djusesikidks ugh ns[krs] rksHkhgstuknZu! dqy ds uk’k ds mRiUunks”kdkstkuusokysgeyksxksbliki ls gVus ds fy;sD;ks ugh fopkjdjukpkfg;sA
10	dqy{k;s iz.k’;fUrdqy/kekZ% lukruk%A /kesZu”Vsdqyad`RLue/keksZ·fHkHkoR;q rAA	dqy ds uk’k ls lukrudqy&/keZu”VgkstkrsagSa] /keZ ds uk’kgkstkusijlEiw.kZdqy es ikiHkhcgqrQSytkrkgSA
11	v/kekZfHkHkokRd`”.kiznq”;fUrdqyfL=;% A L=h”kqnq”Vklq ok”.ksZ; tk;rso.kZlM+dj%AA	gsd`”.k! iki ds vf/kd c<+ tkus ls dqy dh fL=;kavR;Urnwf”krgkstkrhgSvkSjok”.ksZ;! fL+=;ksa ds nwf”krgkstkusijo.kZladjmRiUugksrkgSA

<table>
<tr><td>12</td><td>lM+djksujdk;Sodqy?ukukadqyL; pA
irfUrfirjksás”kkayqIrfi.MksndfØ;k% AA</td><td>o.kZladjdy?kkfr;ksadksdqydksujdystkus ds fy;sgksrkgS A
yqIrgqbZfi.MvkSj ty dh fØ;kokysvFkkZr~ Jk) vkSjriz.koafprbldsfirjyksxHkh
v/kksxfrdksizkIrgksrsgSA</td></tr>
<tr><td>13</td><td>nks”kSjsrS%
dqy?ukukao.kZlM+djdkjdS%A
mRlk|Urstkfr/kekZ% dqy/kekZÜp
‘kkÜork%AA</td><td>blo.kZladjdkjdnks”kksa ls dqy?kkfr;ksa ds
lukrudqy&/keZvkSjtkfr&/keZu”V
gkstkrsgSaA</td></tr>
<tr><td>14</td><td>mRl=dqy/kekZ.kkaeuq”;k.kkatuknZuA
ujds·fu;raoklksHkorhR;uq’kqJqeAA</td><td>gstukjnZu !ftudkdqy /keZu”Vgksx;kgS] ,slseuq”;ksadkvfuf’pedkyrd
ukjhesaoklgksrkgS] ,slkgelqursvk;sgSaA</td></tr>
<tr><td>15</td><td>vgkscregRikiadqrZaO;oflrko;e~A
;nzkT;lq[kyksHksugUrqaLotueq|rkAA</td><td>gka! ‘kksd! geyksxcqf)ekugksdjHkhegku~ ikidjusdksrS;kjgksx;sgS] tks
jkT; vkSjlq[k ds ykHk ls Lotuksadksekjus ds fy;sm|rgksx;sgSA</td></tr>
<tr><td>16</td><td>;fnekeizrhdkje’kL=a ‘kL=ik.k;%A
/kkrZjk”Vªkj.ksgU;qLrUes
{kserjaHkosr~AA</td><td>;fneq> ‘kjfgr ,aolkeuk u djusokysdksgkFkesafy;sgq, /k`rjk”Vª ds iq+=
j.kesaekjMkysrksogekjukHkhesjsfy;svf/kdDy;k.kdkjdgksxkA</td></tr>
<tr><td>17</td><td>,oesqDRoktqZu% lM~[;s
jFkksiLFkmikfo’kr~A
fol`T; l’kjapkia ‘kksdlafoXuekul%AA</td><td>lat; cksys&j.kHkwfeeas ‘kksd ds mf)Xu euokysvtqZublizdkjdgdjok.k
lfgr /kuq”kdksR;kxjdjFk ds fiNysHkkxescSBx;sA</td></tr>
</table>

4.4 Performance Metrics

To evaluate the performance of this proposed corpus based approach with deep neural network, BLEU Score and Word Error Rate is calculated in this work.

- *BLEU Score*

Bleu score is an important metric used for calculating the accuracy of translated sentences as compared to the human generated reference translations. It is not good for shorter translations but it provides accurate results for longer sentences. Normally Bleu Score values lies between 0 and 1, simply multiplying it to 100, its percentage can be calculated. It is observed that higher the bleu score value, model is more accurate. Formula of Bleu Score is as follows:

$$BLEU = \min\left(1, \frac{Output-length}{reference-le}\right)\left(\prod_{i=1}^{4} precision_i\right) \ldots\ldots..(i)$$

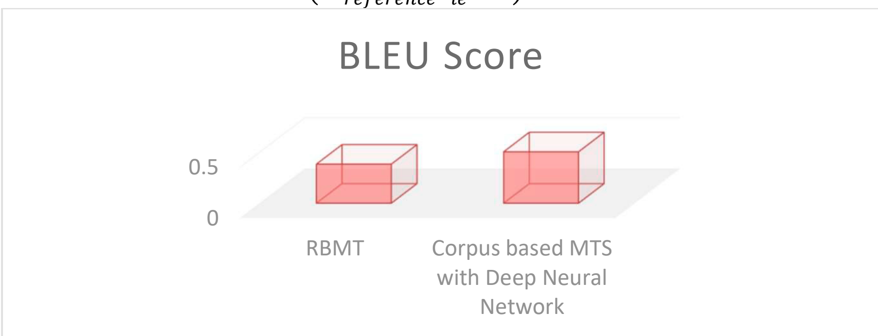


Fig 4: BLEU Score of Rule Based and Proposed MTS

Fig 4 shows that this proposed machine translation system achieves better BLEU Score as compare to rule based machine translation systems. As per results when training of corpus is done with deep neural network then the performance of the corpus-based MTS is 24% better than RBMT.

- *Word Error Rate*

It is a metric used to calculate the error rate by comparing machine translated output with the human translated output. It is assumed that if less the WER, better the model will be

$$WER = \frac{substitutions+insertions+deletion}{reference-len} \quad \text{.........(ii)}$$

Here substitution means replacement of one word with another. Insertion means addition of words and deletion means dropping of words.

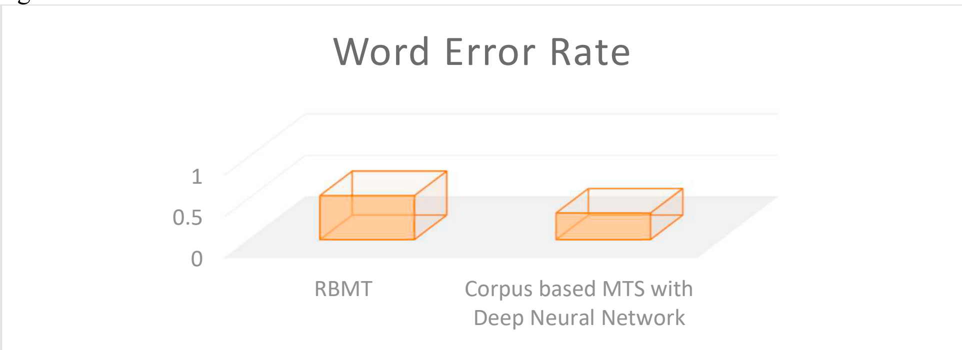


Fig 5: WER of Rule Based and Proposed MTS

Fig 5 shows that the proposed machine translation system has less word error rate which means its performance of the proposed MTS is better than Rule based Machine Translation System by 39.6%.

5. CONCLUSION

Sanskrit to Hindi Translation is one of the most challenging tasks. As a result model became complex and time consuming. In this paper, to overcome existing problem, deep learning concept is used to train the data and for model building. Tensor flow library is used to build a model. In this proposed work, Keras is used as a front end and Tensor flow is used as a back end library. Performance evaluation show that proposed MTS system gives better performance in terms of BLEU Score and WER.

REFERENCES

[1] N. Tapaswi, S. Jain, and V. Chourey, (2012) "Parsing Sanskrit Sentences using Lexical Functional Grammar", International Conference on Systems and Informatics, pp. 2636-2640.
[2] N. Tapaswi, and S. Jain, (2011) "Morphological and Lexical Analysis of the Sanskrit Sentences" MIT International Journal of Computer Science & Information Technology, Vol. 1, No. 1, pp.28-31.
[3] V. M. Barkade, and P. R. Devale,(2010) "English to Sanskrit Machine Translation Semantic Mapper", International Journal of Sciemce and Technology, Vol. 2, Issue 10, pp. 5313-5318.
[4] P. Bahadur, A.k.jain, and D.s.chauhan,(2012) "EtranS- A Complete Framework for English To Sanskrit Machine Translation," International Journal of Advanced Computer Science and Applications, vol. 2, no. 1, pp. 7–13.
[5] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada, (2004) "Fast and optimal decoding for machine translation," Artificial Intelligence, vol. 154, no. 1-2, pp. 127–143.
[6] T. Xiao, J. Zhu, and T. Liu, (2013) "Bagging and Boosting statistical machine translation systems," Artificial Intelligence, vol. 195, no. 6, pp. 496–527.
[7] R. U. U. and T. A. Faruquie, (2005) "An English-Hindi Statistical Machine Translation System," Natural Language Processing – IJCNLP 2004 Lecture Notes in Computer Science, vol. 5, no. 1, pp. 254–262.
[8] Y. Zhang,(2017) "Research on English machine translation system based on the internet," International Journal of Speech Technology, vol. 20, no. 4, pp. 1017–1022.
[9] C. Adak, (2014) "A bilingual machine translation system: English & Bengali," 2014 First International Conference on Automation, Control, Energy and Systems (ACES), pp. 1–4.
[10] L. Chirong, (2014) "Research and Implementation on Machine Translation System with Online Corpora Extraction Technology," Fifth International Conference on Intelligent Systems Design and Engineering Applications, pp. 759–763.
[11] V. Mishra, and R. B. Mishra, (2008) "Study of Example Based English to Sanskrit Machine Translation", Polibits Journal, Vol. 37, pp. 43-54.
[12] S. Kharb, H. Kumar, M. Kumar, and A. K. Chaturvedi, (2017) "Efficiency of a machine translation system," International conference of Electronics, Communication and Aerospace Technology (ICECA), pp. 140–148.

[13] J. Nair, K. A. Krishnan, and R. Deetha, (2016) "An efficient English to Hindi machine translation system using hybrid mechanism," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2109–2113.
[14] B. N. Raju and M. B. Raju, (2016) "Statistical Machine Translation System for Indian Languages,"IEEE 6th International Conference on Advanced Computing (IACC), pp. 174–177.
[15] S. Saini and V. Sahula, (2015) "A Survey of Machine Translation Techniques and Systems for Indian Languages," IEEE International Conference on Computational Intelligence & Communication Technology, pp. 676–681.
[16] R. Östling, Y. Scherrer, J. Tiedemann, G. Tang, and T. Nieminen, (2017) "The Helsinki Neural Machine Translation System," Proceedings of the Second Conference on Machine Translation, pp. 338–347.
[17] D. Ze-Ya, Z. Han-Fen, Z. Quan, M. Jian-Ming, and C. Yu-Huan, (2009) "Automatic Machine Translation Evaluation Based on Sentence Structure Information,"International Conference on Asian Language Processing, pp. 162–166.
[18] R. K. Pandey and G. N. Jha, (2016) "Error Analysis of SaHiT - A Statistical Sanskrit-Hindi Translator," Procedia Computer Science, vol. 96, no. 6, pp. 495–501.
[19] P. Goyal and R. M. K. Sinha, (2008) "Translation Divergence in English-Sanskrit-Hindi Language Pairs," Lecture Notes in Computer Science Sanskrit Computational Linguistics, vol. 8, no. 1, pp. 134–143.
[20] P. Shukla and A. Shukla, (2013) "A Framework of Translator From English Speech To Sanskrit Text ," International Journal of Emerging Technology and Advanced Engineering , vol. 3, no. 11, pp. 113–121.
[21] A. Godase and S. Govilkar, (2015) "Machine Translation Development for Indian Languages and its Approaches," International Journal on Natural Language Computing, vol. 4, no. 2, pp. 55–74, 2015.
[22] N. Sadana, (2017) "Comparison Of Sanskrit Machine Translation Systems," International Journal of Advanced Research in Computer Science, vol. 8, no. 8, pp. 223–225.
[23] J. K. and J. R., (2016) "Sanskrit Machine Translation Systems: A Comparative Analysis," International Journal of Computer Applications, vol. 136, no. 1, pp. 1–4.
[24] V. Mishra and R. Mishra, (2012) "English to Sanskrit machine translation system: a rule-based approach," International Journal of Advanced Intelligence Paradigms, vol. 4, no. 2, pp. 168–184.
[25] J. K. Raulji and J. R. Saini, (2017) "Generating Stopword List for Sanskrit Language," 2017 IEEE 7th International Advance Computing Conference (IACC), pp. 799–802.
[26] S. J, J. S, and D. K. R.N, (2014) "An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism J," International Journal of Engineering and Technology, vol. 6, no. 4, pp. 1909–1919.
[27] H.M. Parmar,(2015) "A Toolkit for Sanskrit Processing".
[28] S. Saini and V. Sahula, (2015) "A Survey of Machine Translation Techniques and Systems for Indian Languages," IEEE International Conference on Computational Intelligence & Communication Technology, pp. 286–290.
[29] Shukla, P., Kulkarni, A., and Shukl, D., (2013) "Geeta: Gold Standard Annotated Data, Analysis and its Application" , ICON.