MACHINE LEARNING BASED TRANSLITERATION FOR MARATHI LANGUAGE USING WER EVALUATION METRICS

[*]Gajanand A. Boywar and Sachin N. Deshmukh
Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS), India - 431004
gaboywar@gmail.com and sndeshmukh@hotmail.com

ABSTRACT

Objective: *To create a Marathi to English transliteration Dataset.*

Methods: *In this paper we tried to explore the WER (Word Error Rate) score evaluation metrics. Using Pair-n Gram model, and LSTM (Long Short Term) sequence to sequence and transformer*

Findings: *Transliteration from the Marathi Script to the English script plays a very important role as Marathi is the official language of Maharashtra and there is a lot of data is present in Marathi which needs to convert into English for global usage and improve the ransliteration rate.*

Novelty: *In the proposed article, we apply the WER metrics and obtain a 1.0 WER score. Machine transliteration is the primary usage of the Dakshina dataset. As a result, we are eager to execute some measures for evaluating transliteration using this dataset as a guide.*

Keywords: *Transliteration, Pair-n gram model, LSTM sequence to sequence model, Transformer sequence to sequence model.*

DOI: 10.48047/ecb/2023.12.6.274

I. INTRODUCTION

English is the largest number of speakers in the world. It is the most widely learned second language. Also in an India. India is the largest democratic country with 22 scheduled languages and 720 dialects used for communication by Indians. Marathi has the third largest number of native speakers in India, after Hindi and Bengali. At 93 million speakers, Marathi is 10th on the world's list of most spoken languages. The most important requirement of translation and transliteration is sharing information, new ideas, thoughts, and facts that are helpful for society.

In today's time, global interaction is increasing day by day and communications between different nationals are done in different languages as well no person knows all the languages. Although English is a global language, not everyone understands it and not every document is available in English. To overcome this barrier of language translation is one very important tool. Converting the author's intended meaning into another language in which they are not fluent it's called translation.

In the translation process, the translator does not use an identical in both the source and target language, it is the meaning which is the important element. Even the sound of the word is usually entirely different. For example, "Hello" in English sounds completely different in Marathi i.e. "नमस्कार". Transliteration is a process of expression of the sound of how a word is pronounced in the source language in the alphabet of the target language. The popular use of it, with names of people or places. For example, "My name is Ram" transliteration becomes "माय नेम इज राम". It's a transliteration of word to word without changing its sound.

A Machine translation system is an automatic system for translating text from one language to another language without human intervention. They play an important role in the field of entertainment, sports, education, offices, tourism, communication, medicine, information technology, research, etc. A few real-time examples where machine translation plays a very important role are cross-lingual question answering, multilingual chat sessions, talking translation applications, and e-mail and website translations. The above starts are just a few of the modern applications of the commercial world.

I.I. MOTIVATION

It is observed that several research works are going on across the globe in their native language to communicate and share ideas across the globe. Even in India, similar work is going on in other languages such as Hindi, Bangla, and Tamil. They have a lot of data in the sense of electronic content. Hindi has the most data set to work, even Bengali also. Hence we will be going to create a similar or as possible new translation model for the Marathi language. When we survey on Marathi data set there is less work done which is not enough for our research, this thing is to motivate me to work on it.

RELATED WORK

As the number of vowels and constants is not the same in all the languages and their corresponding phonemes also are different, one cannot use character matching directly for transliteration. Not all languages have the same sounds/ phonemes for their characters. These missing sounds in a language are created by digraph (two characters) or trigraph (three characters) i.e. by combining two characters or three characters of the language as shown below [Table I].

Table I: Example of digraph and trigraph

Sounds/ phonemes of Marathi Characters not present in English	Equivalent English Character
श	Sh (digraph)
च	Ch (digraph)
क्ष	Ksh (trigraph)

Missing sounds in some languages' pronunciation also create difficulties in transliteration. In the English language, there are several words whose pronunciation is silent. When these languages use words with some silent characters, it becomes difficult to judge which pronunciation technique to use. So the origin of the word is an important aspect to be kept in view for transliteration. Sometimes in one language, a single character represents a specific sound but the same character transliteration in other languages may represent more than one sound.

Sometimes phoneme of the character changes depending on its surrounding characters. The character or set of characters is pronounced differently depending on the words with which these are used.

For the alignment, there are two approaches: Grapheme-based, and Phoneme based.

The Grapheme-based approach - defines the relation and correspondence between the grapheme of the source and target scripts. Different methods are used for the alignment of the grapheme for the character of the source script with the grapheme of the target script. Y. Iia et al. used transliteration as a Statistical Machine Transliteration problem. They used the Noisy channel model for grapheme-based machine transliteration for English to Chinese machine transliteration

The Phoneme-based approach - this approach defines the related correspondence between the phonemes of the source and target script. An alignment of the phoneme for the characters of the source script to the phoneme of the target script is done using different methods.

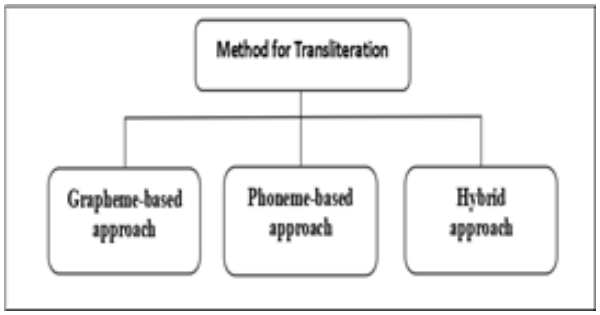

Figure 1: Methods of Transliteration

SMT use phrase translation which creates a phrase translation model and decoder that work with any phrase table. By using the Bayes rule to reformulate translation probability for translating a foreign sentence f (French) into e (English). The decoder develops for comparing different phrased-based translation models employing a beam research algorithm. Partial translation (Hypothesis) is with untranslated foreign words in

the output of the research using the GIZA++ toolkit. Finally, they create a framework (translation model and decoder) that enables us to evaluate and compare various phrase translation methods. The author performs experiments on English-Germany, French-English, English-French, Swedish-English, and Chinese-English language translation. These experiments perform only a few sentences that contain up to three words. The drawback is syntactic model fails to account for important phrase alignments [1].

Statistical phrase-based translation model using hierarchical phrases-phrase that contain sub-phrases. Basic use is synchronous context-free grammar, but at the same time learning from bi-text without any syntactic information. BLEU was used as a metric model improvement of 7.5% over Pharaoh [2], a state-of-art phase-based system. Python technique is used to implement a decoder in this existing system. This author's existing system concludes that the existing system performed without training data of any syntactically annotated base. Also, facilitate the incorporation of syntactic information which did not provide a statically significant gain. The maximum initial phrase length is currently 10, they did not perform any syntactical or grammar-based translation [2].

English to French translation tasks shows that this method provides sub-static improvement of up to 2.8 BLEU (bilingual language understudy) points. NMT is in-capable of rare words translated because they have fixed modest-sized vocabulary which forces them to use *unk* symbols. The authors create a simple alignment-based technique to overcome rare words problem using NMT than any other technique, just like LSTM. The existing system gets 2.8 BLEU points over various NMT systems, most important this was created in the WMT'14 contest dataset which states as best MT at that time. For this existing system, the dataset size is 12M parallel sentences (348M French and 304M English words) [3].

Continuous Translation Models are also stated as Recurrent Continuous Translation Models which are purely based on the continuous representation of words, phrases, and sentences. Existing models obtain difficulty concerning gold translation that is > 43% lower than that of state-of-art alignment-based translation models. It's sensitive to word order, syntax, and meaning of the source sentences despite lacking alignment.

When recurring n-best lists of translation they match the state-of-art system. While performing this experiment the author uses RCTM, RLM, CSM RCTM II, and conventional n-gram models. The size of a dataset for experiments used a bilingual corpus of 144953 pairs of sentences which contains less than 80 words in a length of new commentary. Proposed model capture syntactic and semantic information and estimates during re-ranking the quality of candidate translations. An RCTM offers flexibility due to its sensitivity to the continuous representation of conditioning information. It suggests a wide range of potential for single sentences, and multilingual source representation can also able to do with better improved by this model [4].

DEEP Neural Networks (DNNs) are powerful models that have excellent performance on difficult learning tasks. DNN works at large training datasets available, without mapping sequences to sequences. The aim of the author is an end-to-end approach, so the sequence will make minimum assumptions on the structure of the sequence. They used Multi-layered LSTM to map the source sentences to a vector of a fixed multidimensional and another LSTM to decode. LSTM achieve a BLEU score of 34.8 out of the entire set in the WMT'14 dataset. LSTM did not have to face any difficulty with long sentences. 1000 hypotheses dataset produces by the SMT system, and its score increased to 36.5 which is tremendously increased. Using dependencies between source sentences to target sentences helps to increase after reversing the order of the words in all source sentences and it improved the LSTM's performance markedly.

The existing system worked on a trained subset of 12M sentences consisting of 348M French words and 304M English words. The result is improved by reversing the words in the source sentences. It's important to find problem encoding that has the greatest number of short-term dependencies, as they make the learning problem much simpler. LSTM translate very long sentence correctly. Maybe standard RNN should be easily trainable when the source sentences are reversed [5].

The ROVER approach describes the calculation of consensus transformations from the output of multiple machine translation (MT) systems. Combined outputs and generates a new hypothesis. The ROVER approach is used for speech recognition hypotheses to create pairwise word alignments for the original MT hypothesis using a statistical alignment algorithm that explicitly models word rearrangement. The whole document gets translated rather than single sentences taken to produce alignment. BLEU (bilingual language understudy) score by 15% [6].

The Mellow [7] Multi-engine machine merges the output from several machine translation systems into a single translation. Combined Arabic-English output scored 5.22 BLEU points higher than the best individual system. The source code of the system significantly improved on some translation tasks, only closed in the performance system. The software can be downloaded, installed, and run. The ARPA model and METEOR software are needed if necessary. This software was developed using the C++ compiler, Java, and Python [7].

The multiple System combination models can exploit phrasal and structural system-weighted consensus and also utilize existing information about word ordering present in the target hypothesis. The sentence-level paraphrasing for the machine transition system combination has been done. They developed a hybrid combination architecture to operate on the hypothesis using different phasing techniques. A significant improvement over combination baselines. Many simple frameworks can borrow the modified for the combination [8].

LSTM extends a tree structure, which shows the history of multiple child cells or multiple descendant cells in a recursive process. The S-LSTM model provides long-distance interaction over hierarchy. Recursion is a fundamental process associated with many problems using Recursive neural networks, recurrent networks, and LSTM. This existing memory cell reflects the history memories of multiple descendants through generated copying of memory vectors. S-LSTM achieves state-of-the-art performance after replacing the recursive model with tensor-enhanced compositional layers with S-LSTM [9].

The combination model implemented targeted Jane, RWTH's open-source statistical machine translation toolkit. Existing system combination pipeline with additional n-gram language models and lexical translation model. In the WMT 2011 submission of Jane, the open-source machine translation model was similar level or best at that time, now in WMT-14 it improve by 0.7 pints in BLEU, so the existing Jane includes State-of-the-Art (SoA), it also improved comparatively than n-gram model and IBM-1 [10].

The researcher paired monolingual data with NMT which can learn the same language information. WMT 15 English to German and German to English achieve 2.8-3.7 BLEU as compared to TWSLT14 task which performs on Turkish to English and English to Turkish which is between 2.1-3.4 BLEU points [10]. Research encodes the rare word and unknown word as a sequence of subword units. By using n-gram models and byte pair encoding, which is used for word segmentation, which encodes the open vocabularies with compact symbol vocabulary of variable-length of subword units. These both compression algorithm which introduced in the WMT-15 translation task for English to German and English to Russian by using vocabulary neural models which contain 30000-50000 words, and the researcher got BLEU point 1.1 and 1.3 respectively [11].

Existing systems work on rare words which never exist in bilingual training data. They introduced two methods one is to mix the word-character model to analyze other is to check parallel sentence which gives a similar translation of the lexicon. While creating a bridge between NMT and Bilingual dictionary it takes a dataset of 630K Chinese-English bilingual sentences, 100M Chinese sentences, and 86252 translation lexicons are used. Re-labeling the OOV words with characters sequences over the problem of the word/characters model. 70% of rare or unseen words get correctly translated. The data synthesis model does not distinguish the original training data [12]. The aim is to maximize the probability of an English sentence given French and German sources. Using neural encoder-decoder framework. By applying this method they create a multi-source machine translation model [13], which helps to disambiguate the word sense [13].

NMT significantly translated sentences are more accurate and fluent than translation by SMT-based system. Sometimes NMT produced a different meaning this happens only when rare words occur. The author uses phrase-based machine translation to pre-translate the input into a target language. Then NMT generates hypotheses using pre-translation. They create techniques for the English-to-German translation task. After combining NMT and SMT as a PBMT it more easily translates the sentence, and also translate rare word [12]. SBSC can't detect all features, this problem solves by a neural-based approach for system combination. NBSC gains a maximum accuracy of features [14].

The researcher's existing method follows a combination of Chinese segmentation of different results, such as words, their alignment, and adding reliable bilingual words with a high probability of training data, which contain person name, locations name, organization names, temporal and numerical expression. Replacing Chinese characters with Chinese pinyin for purpose of training translation model from Chinese to English. The researcher follows three decoder techniques (i) Moses - which shows the latest development in the area of statistical machine translation research, (ii) Joshua- a hierarchical phased statistical machine translation decoder, that easily runs on large-scale data, (iii)MEBTG - a maximum Entropy-based reordering model

decoder, which is the prediction of relative orders of any two adjacent blocks contain a problem of classification[18].

Researchers aim to develop a system model which is the combination of multiple models which work in three-layer to achieve better translation output. For a combination of models, they used SMT and NMT models. The advantage of these models is they give increased accuracy of the translation. For this they used the HindiEnCorp corpus, containing 273880 sentences trained datasets are used for SMT, NMT, and NBSC models. After that, they used a phrase-based statistical machine translation (PBMT) system, hierarchical phrase-based System (Hiero), NMT, and Google translation engine. Then researcher create an HMEMT (Hybrid Multi-Engine Machine Translation) Model, and it gives tremendously improved results with a BLEU point of 19.97. Which is for English to Hindi and Hindi to English, which is a 95% confidence level [19].

As per mention in [**Error! Reference source not found.**] and a review, the overall outcome is that the hybrid approach can be the solution for the translation and transliteration of any native to non-native languages. It varies on the researcher which kind of hybrid approach they create either SMT+NMT or RBMT+SMT or a newer one.

II. Data and Evaluation Metrics

We used information from Marathi language Wikipedia material that was extracted in March 2019.

After collecting material from Wikipedia, we pre-processed the text using several methods. Then, for comparative analysis, we use a Dakshina dataset as a reference. Then, upon the completion of a BLEU evaluation, measurement metrics.

The Evaluation has been based on the Levenshtein distance algorithm which is used to calculate the WER (Word Error Rate) the formula for the calculate error rate is as shown in below formula (i)

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad \text{.....(i)}$$

Where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of Corrects,
- N is the number of words in the reference, N=(S+D+C)

III. RESULT

We used WER score evaluation metrics, and we got the 1.0 WER , Number of Substitute character is 1 as shown in Figure 2: Result of WER (Word Error Rate)Figure 2

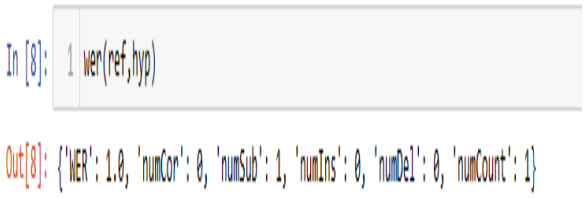

Figure 2: Result of WER (Word Error Rate)

IV. CONCLUSION

Using the Dakshina dataset as reference, we have calculated WER score for Marathi text, which is 1.0. The goal of our feature set is to improve the outcome in comparison to competitors. Additionally, Dakshina dataset is being used for Marathi text transliteration. We anticipate using this dataset to facilitate experimentation on transliteration.

ACKNOWLEDGMENTS

Authors would like to acknowledge thanks to Dr. Babasaheb Ambedkar Research and Training Institute, Pune (BARTI) for financial support in this work. The authors also would like to thank the Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS), India for providing infrastructure and necessary support for carryout the research.

REFERENCES

[1] B. Bangalore, G. Bordel and G. Riccardi, "Computing consensus translation from multiple machine translation systems," *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU* '01., 2001, pp. 351-354, https://doi.org/10.1109/ASRU.2001.1034659

[2] Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. *In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133. https://aclanthology.org/N03-1017

[3] Chiang D. A hierarchical phrase-based model for statistical machine translation. InProceedings of the 43rd annual meeting of the association for computational linguistics (acl'05) 2005 Jun (pp. 263-270). https://aclanthology.org/P05-1033

[4] Li M, Zhang J, Zhou Y, Zong C. The CASIA statistical machine translation system for IWSLT 2009. *InProceedings of the 6th International Workshop on Spoken Language Translation: Evaluation Campaign 2009.* https://aclanthology.org/2009.iwslt-evaluation.13.pdf

[5] Heafield K, Lavie A. Combining machine translation output with open source. *The carnegie mellon multi-engine machine translation scheme. The Prague Bulletin of Mathematical Linguistics.* 2010 Jan 26;93(1):27-36. https://aclanthology.org/www.mt-archive.info/10/MTMarathon-2010-Heafield.pdf

[6] Barrault L. MANY: Open source machine translation system combination. *The Prague Bulletin of Mathematical Linguistics.* 2010;93:147. https://doi.org/10.2478/v10108-010-0001-y

[7] Kalchbrenner N, Blunsom P. Recurrent continuous translation models. *InProceedings of the 2013 conference on empirical methods in natural language processing 2013* Oct (pp. 1700-1709). https://aclanthology.org/D13-1176.pdf

[8] Luong MT, Sutskever I, Le QV, Vinyals O, Zaremba W. Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206. 2014 Oct 30. https://doi.org/10.48550/arXiv.1410.8206

[9] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Advances in neural information processing systems. 2014;27. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[10] Freitag M, Huck M, Ney H. Jane: Open source machine translation system combination. *InProceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics 2014 Apr* (pp. 29-32). https://aclanthology.org/E14-2008.pdf

[11] Ma WY, McKeown K. System combination for machine translation through paraphrasing. *InProceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 2015 Sep* (pp. 1053-1058). https://aclanthology.org/D15-1122.pdf

[12] Zhu X, Sobhani P, Guo H. Long short-term memory over tree structures. arXiv preprint arXiv:1503.04881. 2015 Mar 16. https://doi.org/10.48550/arXiv.1503.04881

[13] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709. 2015 Nov 20. https://doi.org/10.48550/arXiv.1511.06709

[14] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909. 2015 Aug 31. https://doi.org/10.48550/arXiv.1508.07909

[15] Zhang J, Zong C. Bridging neural machine translation and bilingual dictionaries. arXiv preprint arXiv:1610.07272. 2016 Oct 24. https://doi.org/10.48550/arXiv.1610.07272

[16] Zoph B, Knight K. Multi-source neural translation. arXiv preprint arXiv:1601.00710. 2016 Jan 5. https://doi.org/10.48550/arXiv.1601.00710

[17] Niehues J, Cho E, Ha TL, Waibel A. Pre-translation for neural machine translation. arXiv preprint arXiv:1610.05243. 2016 Oct 17. https://doi.org/10.48550/arXiv.1610.05243

[18] Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural System Combination for Machine Translation. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–384, Vancouver, Canada. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P17-2060

[19] Banik D, Ekbal A, Bhattacharyya P, Bhattacharyya S. Assembling translations from multi-engine machine translation outputs. *Applied Soft Computing.* 2019 May 1;78:230-9. https://doi.org/10.1016/j.asoc.2019.02.031

[20] Roark B, Wolf-Sonkin L, Kirov C, Mielke SJ, Johny C, Demirsahin I, Hall K. Processing south asian languages written in the latin script: the dakshina dataset. arXiv preprint arXiv:2007.01176. 2020 Jul 2. https://doi.org/10.48550/arXiv.2007.01176